Survival of packages on CRAN

Xuan Fang

06/2020

Contents

Background

There are a lot of R packages on CRAN, with the number growing exponentially over time. However, not all the packages stay there: they may be removed from the live version of CRAN by the author's request or because they don't work with the current version of R. All current and past packages are stored in the CRAN archive.

In recent years, many more people have been writing packages, and packages have tended to have more *dependencies* on other packages. This might be expected to lead to packages that don't last as long.

The file `alldates.rda` contains a single data frame `all_dates`, with variables listed as follows:

Variable	Description
pkg	the name of the package
cran_date	the date the current version was put on CRAN
first	the oldest date when a version was put into archive
latest	the most recent date when a version was put into archive

Note: The dates are offsets from 1970/1/1

The file `snapshots.rda` contains two data frames:

a: the result of `available.packages()`, a matrix listing all available CRAN packages on 2020-6-2.

b: the result of `available.packages()` as it would have been on 2015-6-2, via the Microsoft CRAN Time Machine.

Our Tasks are:

1. Describe how the longevity of CRAN packages has changed over time
2. Describe how version number, dependencies, and license relate to the probability of a 2015 package surviving to 2020 (as a binary outcome)
3. Describe how dependencies, and license relate to the subsequent survival time of a package on CRAN in 2015

Load datasets

```
# Load the data
load(file = "alldates-1.rda")
load(file = "snapshots.rda")
```

Task One

Section 1: Data exploratory and data cleaning

Pre-analysis exploratory: Examining `all_dates` dataset

```
str(all_dates)
```

```
## 'data.frame':    18652 obs. of  4 variables:
##  $ pkg      : Factor w/ 18652 levels "A3","aaSEA","ABACUS",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ cran_date: num  16663 18209 18159 18072 16560 ...
##  $ first    : num  15743 18109 NA 16598 14887 ...
##  $ latest   : num  15791 18109 NA 17682 16263 ...
```

We notice there are 'Inf'/'-Inf' values in the `first` and `latest` columns.

```
summary(all_dates)
```

```
##        pkg          cran_date         first           latest
##  A3      :    1   Min.   :13222   Min.   :10143   Min.   : -Inf
##  aaSEA   :    1   1st Qu.:17269   1st Qu.:15427   1st Qu.:16495
##  ABACUS  :    1   Median :17961   Median :16645   Median :17475
##  abbyyR  :    1   Mean   :17672   Mean   :  Inf   Mean   : -Inf
##  abc     :    1   3rd Qu.:18298   3rd Qu.:17428   3rd Qu.:18024
##  abc.data:    1   Max.   :18414   Max.   :  Inf   Max.   :18412
##  (Other) :18646   NA's   :2939    NA's   :3757    NA's   :3757
```

```
head(all_dates)
```

```
##        pkg cran_date first latest
## 1       A3     16663 15743  15791
## 2    aaSEA     18209 18109  18109
## 3   ABACUS     18159    NA     NA
## 4   abbyyR      18072 16598  17682
## 5      abc      16560 14887  16263
## 6 abc.data      16560    NA     NA
```

Data cleaning

Remove records that have `first` = inf (and `latest` = -inf), which are implausible. But for `pkg = "sergeant"`, we set its `first` and `latest` columns to `NA` since it has value in `cran_date`.

```r
# Have a look at the records that have `first = inf` (and `latest = -inf`):
all_dates[is.infinite(all_dates$first), ]
```

```
##                   pkg cran_date first latest
## 12885        sergeant     18414   Inf   -Inf
## 15714      2020-05-14        NA   Inf   -Inf
## 15763             and        NA   Inf   -Inf
## 15788        Archived        NA   Inf   -Inf
## 15798              as        NA   Inf   -Inf
## 15879            been        NA   Inf   -Inf
## 16044           check        NA   Inf   -Inf
## 16143       corrected        NA   Inf   -Inf
## 16226          datasa        NA   Inf   -Inf
## 16269       depending        NA   Inf   -Inf
## 16274         despite        NA   Inf   -Inf
## 16641           given        NA   Inf   -Inf
## 16745            have        NA   Inf   -Inf
## 16833              in        NA   Inf   -Inf
## 16880              it        NA   Inf   -Inf
## 17005            long        NA   Inf   -Inf
## 17347         notice.        NA   Inf   -Inf
## 17394              on        NA   Inf   -Inf
## 17420        orphaned        NA   Inf   -Inf
## 17605        problems        NA   Inf   -Inf
## 17811      reminders.        NA   Inf   -Inf
## 18400           those        NA   Inf   -Inf
## 18409           time.        NA   Inf   -Inf
## 18588            were        NA   Inf   -Inf
## 18617 X-CRAN-Comment:        NA   Inf   -Inf
```

```r
# For pkg = "sergeant", we set first and latest to "NA":
all_dates$first[all_dates$pkg=="sergeant"] <- NA
all_dates$latest[all_dates$pkg=="sergeant"] <- NA
```

```r
# For the rest records, we just delete them, and now we get 18,628 records:
all_dates <- subset(all_dates, !is.infinite(all_dates$first))
summary(all_dates)
```

```
##        pkg           cran_date         first            latest     
##  A3      :    1   Min.   :13222   Min.   :10143   Min.   :10143  
##  aaSEA   :    1   1st Qu.:17269   1st Qu.:15423   1st Qu.:16502  
##  ABACUS  :    1   Median :17961   Median :16643   Median :17478  
##  abbyyR  :    1   Mean   :17672   Mean   :16263   Mean   :17151  
##  abc     :    1   3rd Qu.:18298   3rd Qu.:17424   3rd Qu.:18025  
##  abc.data:    1   Max.   :18414   Max.   :18411   Max.   :18412  
##  (Other) :18622   NA's   :2915    NA's   :3758    NA's   :3758   
```

```
# If package is not 'alive', there will have an 'NA' in cran_date column. (2915 records)
summary(all_dates[is.na(all_dates$cran_date), ])
```

```
##             pkg         cran_date         first            latest     
##  aaMI        :   1   Min.   : NA    Min.   :10143   Min.   :10143  
##  ABCExtremes :   1   1st Qu.: NA    1st Qu.:14651   1st Qu.:15525  
##  abemus      :   1   Median : NA    Median :15692   Median :16274  
##  aBioMarVsuit:   1   Mean   :NaN    Mean   :15599   Mean   :16228  
##  accuracy    :   1   3rd Qu.: NA    3rd Qu.:16863   3rd Qu.:17317  
##  Ace         :   1   Max.   : NA    Max.   :18396   Max.   :18396  
##  (Other)     :2909   NA's   :2915                                  
```

Commentary: There are 2,915 records that "cran_date" is 'NA', suggesting those packages are died (i.e. have been removed).

For those records that `cran_date` is not 'NA' (meaning it is alive) but the `first` and `latest` are both 'NA' means the version on CRAN is the first version, so there are no archived versions.

```
# Take a look at those records (3758 records):
test <- subset(all_dates, is.na(first)&is.na(latest))
test$cran <- as.Date(test$cran_date)
summary(test)
```

```
##         pkg         cran_date         first          latest    
##  ABACUS  :   1   Min.   :13222   Min.   : NA    Min.   : NA    
##  abc.data:   1   1st Qu.:16900   1st Qu.: NA    1st Qu.: NA    
##  ABC.RAP :   1   Median :17605   Median : NA    Median : NA    
##  abcADM  :   1   Mean   :17411   Mean   :NaN    Mean   :NaN    
##  abdiv   :   1   3rd Qu.:18110   3rd Qu.: NA    3rd Qu.: NA    
##  abe     :   1   Max.   :18414   Max.   : NA    Max.   : NA    
##  (Other) :3752                   NA's   :3758   NA's   :3758   
##       cran           
##  Min.   :2006-03-15  
##  1st Qu.:2016-04-09  
##  Median :2018-03-15  
##  Mean   :2017-09-01  
##  3rd Qu.:2019-08-02  
##  Max.   :2020-06-01  
##                      
```

Adding new variables:

First, we want to create a new variable `end_date` that is the removal date for packages that are removed, or presumably at the time the data were downloaded, which is 2020-6-2. If the package is still alive, it is censored either when the data was retrieved or at the date the current version appeared on CRAN. There are arguments in favour of either choice. Nevertheless, we decide to set the `end_date` as the date were downloaded.

```
end_date_num <- ifelse(is.na(all_dates$cran_date), all_dates$latest, as.Date("2020-6-2"))
all_dates$end_date <- end_date_num
summary(all_dates)
```

```
##         pkg          cran_date          first            latest     
##  A3      :    1   Min.   :13222   Min.   :10143   Min.   :10143  
##  aaSEA   :    1   1st Qu.:17269   1st Qu.:15423   1st Qu.:16502  
##  ABACUS  :    1   Median :17961   Median :16643   Median :17478  
##  abbyyR  :    1   Mean   :17672   Mean   :16263   Mean   :17151  
##  abc     :    1   3rd Qu.:18298   3rd Qu.:17424   3rd Qu.:18025  
##  abc.data:    1   Max.   :18414   Max.   :18411   Max.   :18412  
##  (Other) :18622   NA's   :2915    NA's   :3758    NA's   :3758   
##     end_date    
##  Min.   :10143  
##  1st Qu.:18415  
##  Median :18415  
##  Mean   :18073  
##  3rd Qu.:18415  
##  Max.   :18415  
## 
```

Then, we create a new variable `removed` as an event(removal) indicator.

```
all_dates$removed<- is.na(all_dates$cran_date)
```

Section 2: Data analysis

The biggest problem is that we don't have the date the package first appeared on CRAN unless the current version is the first version: `first` is the first date the package was archived, which is either the date of the second version or is when the package was removed from CRAN, whichever is earlier. We're going to use `first` as the start date because it's what we've got and it isn't bad.

We're looking at survival from `first` arrival on CRAN, which is time zero, and the end of observation time is `end_date-first`. The event indicator is `is.na(cran_date)`, which we created previously as 'removed".

```
summary(as.Date(all_dates$first))
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max. 
## "1997-10-09" "2012-03-23" "2015-07-27" "2014-07-11" "2017-09-15" "2020-05-29" 
##         NA's 
##       "3758"
```

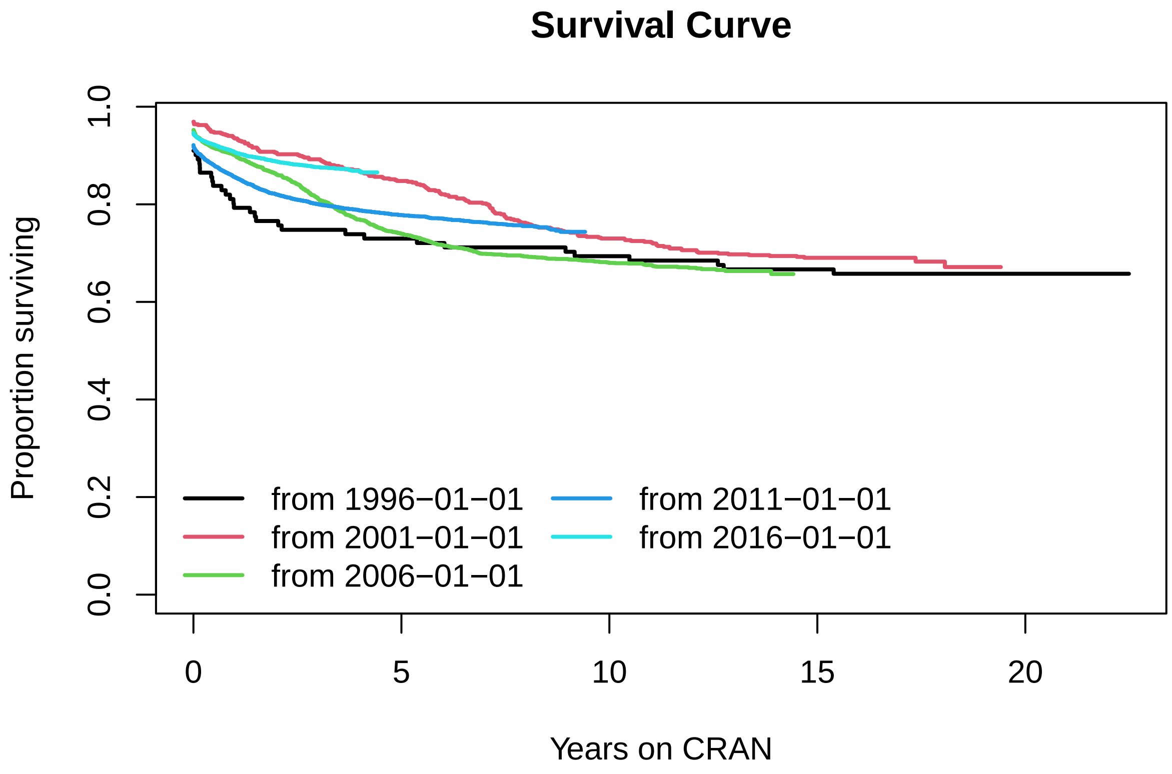
Fit a Cox model by defining a `start_period` variable:

```
cut_points<- as.Date(c("1996-1-1","2001-1-1","2006-1-1","2011-1-1","2016-1-1","2020-6-3"))
all_dates$start_period <- with(all_dates, cut(as.Date(first), cut_points))
coxph(Surv(end_date-first, removed)~start_period, data=all_dates)
```

```
## Call:
## coxph(formula = Surv(end_date - first, removed) ~ start_period, 
##     data = all_dates)
## 
##                              coef exp(coef)  se(coef)      z        p
## start_period2001-01-01 -0.163520  0.849149  0.179085 -0.913 0.361196
## start_period2006-01-01  0.006469  1.006490  0.168472  0.038 0.969368
## start_period2011-01-01 -0.156825  0.854854  0.167068 -0.939 0.347889
## start_period2016-01-01 -0.560433  0.570962  0.169069 -3.315 0.000917
## 
## Likelihood ratio test=122.6  on 4 df, p=< 2.2e-16
## n= 14870, number of events= 2915 
##    (3758 observations deleted due to missingness)
```

Visualization:

```
cran_periods<-survfit(Surv(end_date-first, removed)~start_period, data=all_dates)
plot(cran_periods, col = 1:5, lwd = 2, xscale = 365,
     xlab = "Years on CRAN", ylab = "Proportion surviving", main = "Survival Curve")
legend("bottomleft", bty = "n", ncol = 2, lwd = 2, col = 1:5, legend=paste("from",cut_points)[-6])
```



Summary The hazard ratio for 2001-2010 group of packages to 1996-2000 group of packages (baseline group) is 0.849, so 2001-2010 group of packages had a 15.1% lower rate (hazard) of death than the baseline group (on average, over the follow-up time);

The hazard ratio for 2006-2010 group of packages to 1996-2000 group of packages (baseline group) is 1.006, so 2006-2010 group of packages had a hazard of death very similar to the baseline group (on average, over the follow-up time);

The hazard ratio for 2011-2015 group of packages to 1996-2000 group of packages (baseline group) is 0.855, so 2011-2015 group of packages had a 14.5% lower rate (hazard) of death than the baseline group (on average, over the follow-up time);

The hazard ratio for 2016-2020 group of packages to 1996-2000 group of packages (baseline group) is 0.571, so the 2016-2020 group of packages had a 42.9% lower rate (hazard) of death than the baseline group (on average, over the follow-up time).

In conclusion, early packages died faster, then things improved, but have gotten worse around 2010. One possible explanation is that package namespaces were made compulsory in R 2.14.0. And the hazard rate (survival has been higher) has been lower in recent years, suggesting longer longevity in recent years.

Task Two

Describe how version number, dependencies, and license relate to the probability of a 2015 package surviving to 2020 (as a binary outcome)

Convert matrix to data frame:

```r
# For a (the result of available.packages(), a matrix listing all available CRAN packages on 2020-6-2.)
A.df <- as.data.frame(a, row.names = FALSE)
A.df <- A.df[c("Package", "Version", "Depends", "License")]
head(A.df)
```

```
##    Package Version                                             Depends
## 1       A3   1.0.0                      R (>= 2.15.0), xtable, pbapply
## 2    aaSEA   1.1.0                                         R(>= 3.4.0)
## 3   ABACUS   1.0.0                                        R (>= 3.1.0)
## 4   abbyyR   0.5.5                                        R (>= 3.2.0)
## 5      abc     2.1 R (>= 2.10), abc.data, nnet, quantreg, MASS, locfit
## 6 abc.data     1.0                                         R (>= 2.10)
##              License
## 1         GPL (>= 2)
## 2              GPL-3
## 3              GPL-3
## 4 MIT + file LICENSE
## 5         GPL (>= 3)
## 6         GPL (>= 3)
```

```r
# For b (the result of available.packages() as it would have been on 2015-6-2, via the Microsoft CRAN T
B.df <- as.data.frame(b, row.names = FALSE)
B.df <- B.df[c("Package", "Version", "Depends", "License")]
B.df <- B.df[!is.na(B.df$Package), ]
head(B.df)
```

```
##       Package Version                                             Depends
## 1          A3   0.9.2                      R (>= 2.15.0), xtable, pbapply
## 2         abc     2.1 R (>= 2.10), abc.data, nnet, quantreg, MASS, locfit
## 3 ABCanalysis   1.0.1                                         R (>= 2.10)
## 4    abc.data     1.0                                         R (>= 2.10)
## 5    abcdeFBA     0.4              Rglpk,rgl,corrplot,lattice,R (>= 2.10)
## 6 ABCExtremes     1.0                           SpatialExtremes, combinat
##      License
## 1 GPL (>= 2)
## 2 GPL (>= 3)
## 3      GPL-3
## 4 GPL (>= 3)
## 5      GPL-2
## 6      GPL-2
```

Create new variables

Binary outcome variable: Survived

Create a binary for the 2015 package surviving to 2020:

```r
Survived <-  B.df[,"Package"] %in% A.df[,"Package"]
# Take a look at the result:
table(Survived)
```

```
## Survived
## FALSE  TRUE
##   936  5764
```

```
# Convert it to 0/1 variable: 0 for survived, 1 for died:
B.df$Survived <- as.factor(as.numeric(B.df[,"Package"] %in% A.df[,"Package"]))
```

Version variable: Version_new

Extract the first numeric component (before the first 'dot') of the Version variable and make a new variable Version_new:

```
Version_new <- substr(B.df$Version, 1, regexpr('[.]', B.df$Version)-1)
# Take a look at the result:
table(Version_new)
```

```
## Version_new
##               0       1      14      15       2    2011    2012    2013 2013-10
##       6    2527    3293       1       1     550       3       5       9       1
## 2013-11  2013-2  2013-9    2014    2015  2015-2    2152    2160       3    3000
##       1       1       1      13      25       2       1       1     136       1
##    3010    3011    3012       4       5       6       7       8       9
##       9       6       2      69      19       4       6       6       1
```

We define Version_new that does not fall into (0, 4) as "others":

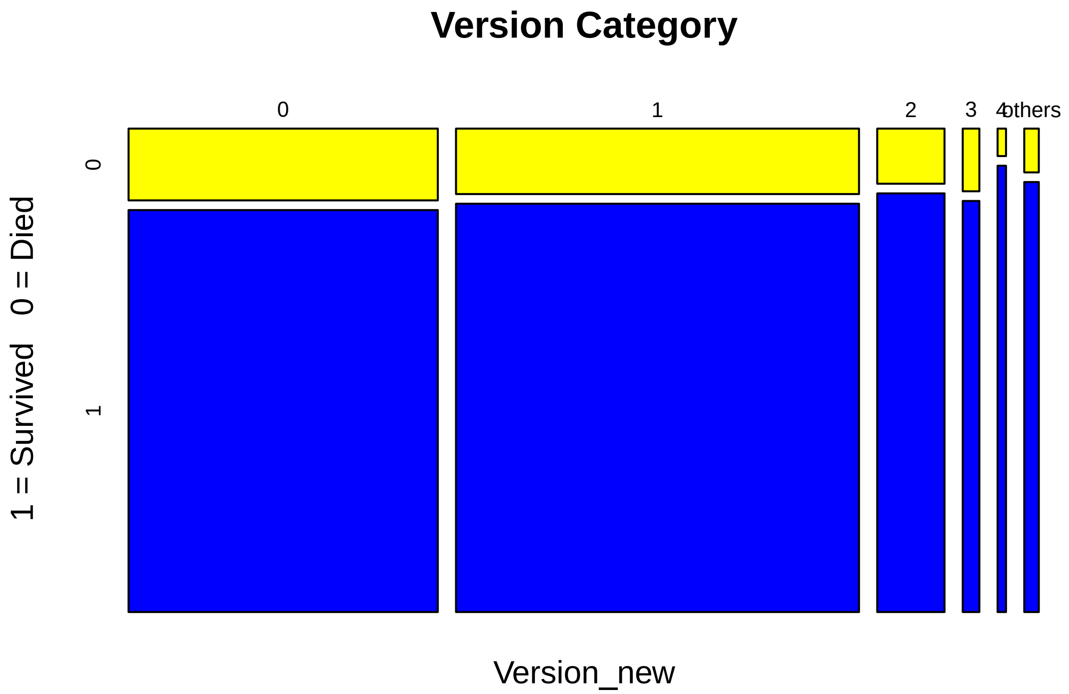
```
B.df$Version_new <- as.factor(ifelse(as.numeric(substring(Version_new, 1, 2))<=4, Version_new, "others")
table(B.df$Version_new)
```

```
##
##      0      1      2      3      4 others
##   2527   3293    550    136     69    119
```

Commentary: In the 2015 data, most of the observations fall into category '1' and '0', then followed by category '2', '3', and '4'. The rest of categories are minority, thus in order to reduce the levels of the Version_new variable we make them into category 'others'.

Visualization:

```
# Mosaic plot of the version category v.s. Survived
with(B.df, mosaicplot(table(Version_new, Survived), col=c("yellow","blue"),
                      main="Version Category", ylab = "1 = Survived   0 = Died"))
```



Commentary: The mosaic plot shows version category '4' has the largest proportion of packages survived, followed by 'others', '2'; version category '1' and '3' has the similar proportion of packages survived; version category '0' has the lowest proportion of survival. It seems that the updates version (which has higher first numeric component) tend to have higher proportion of survival.

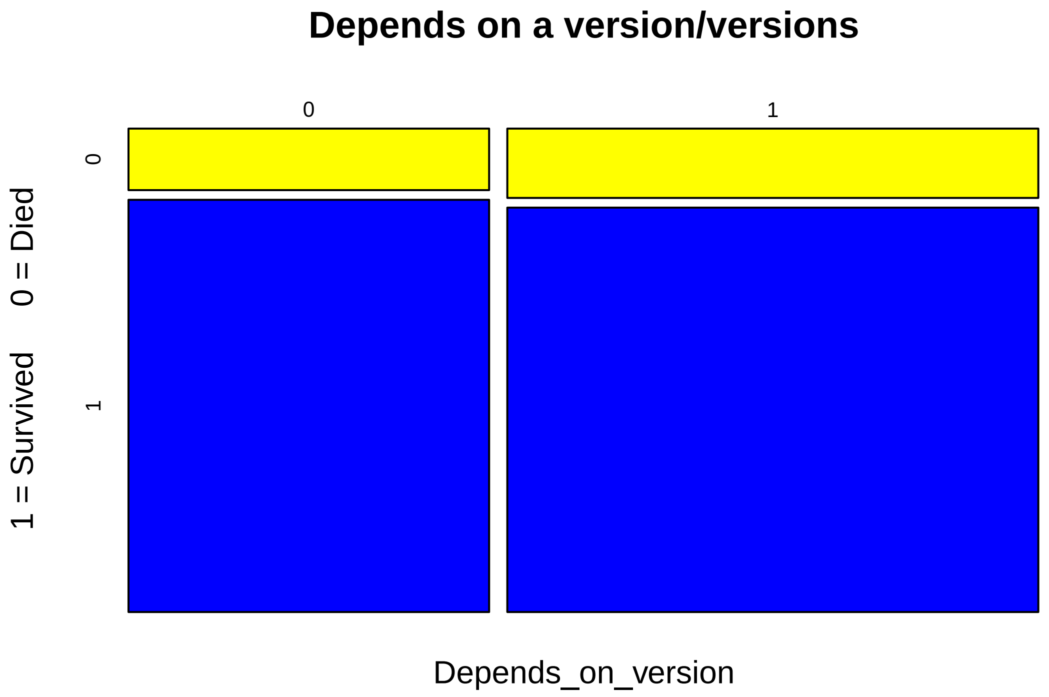
Dependencies variables: `Depends_on_version` and `Num_Depends`

1) Create a binary `Depends_on_version` for dependencies on a version/versions: we know that there will be parentheses in the variable if the package has dependencies on a version/versions:

```
# Convert it to 0/1 variable: 0 for non-dependency, 1 for dependency:
B.df$Depends_on_version <- as.factor(as.numeric(grepl("[(]>=", B.df$Depends)))
```

Visualization:

```
# Mosaic plot of the version  Depends_on_version v.s. Survived
with(B.df, mosaicplot(table(Depends_on_version, Survived), col=c("yellow","blue"),
                      main="Depends on a version/versions", ylab = "1 = Survived     0 = Died"))
```



Commentary: It seems that a package that has no dependency on other a version/versions has a relatively higher probability of survival rate, but that not be significant.

2) Create a variable `Num_Depends` which count the number of dependencies:

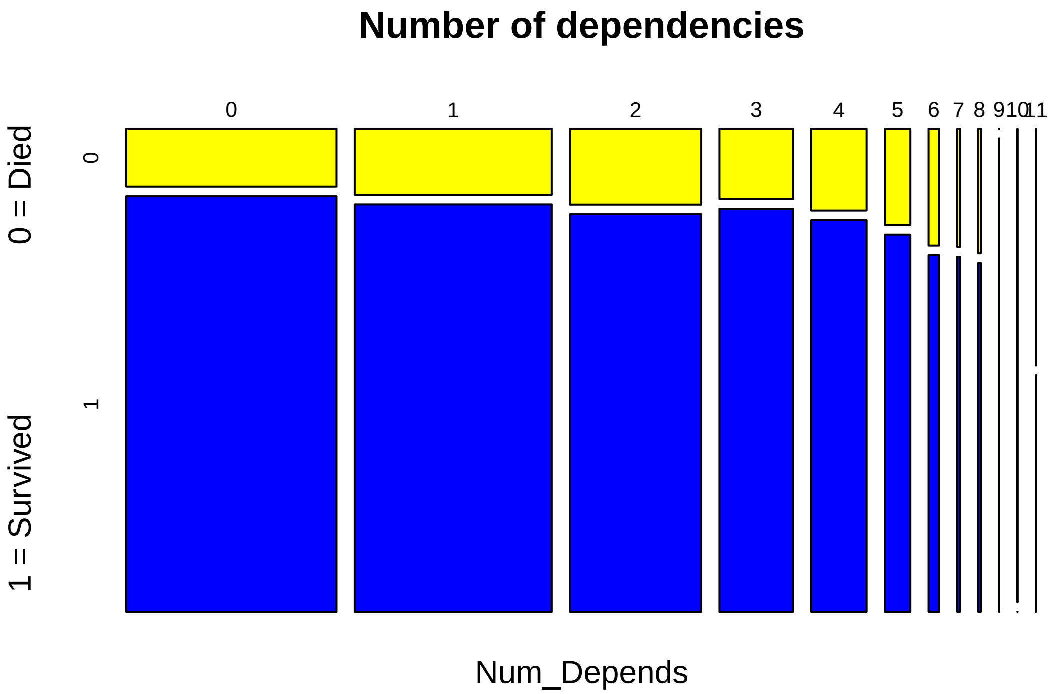
```
# Split the strings:
Depends_split <- strsplit(as.character(B.df$Depends), ",")
# Use gregexpr to count the number of dependences:
B.df$Num_Depends <- sapply((gregexpr("[A-Za-z]{2,}", Depends_split, ignore.case = TRUE)),
                           function(i) sum(i > 0))
```

```
table(B.df$Num_Depends)
```

```
##
##    0    1    2    3    4    5    6    7    8    9   10   11
## 1538 1441  961  538  405  187   77   20   19    1    2    2
```

Visualization:

```
# Mosaic plot of the version Num_Depends v.s. Survived
with(B.df, mosaicplot(table(Num_Depends, Survived), col=c("yellow","blue"),
                      main="Number of dependencies", ylab = "1 = Survived                    0 = Died"))
```



Commentary: The mosaic plot shows that the higher number of dependencies on other packages, the lower the survival probability of a package is.

License variables: `License_new` & `License_alt`

1) Create a new variable `License_new`, which regroup the license categories.

```
# Take a look at the frequency table of License
table(B.df$License)
```

```
## 
##                            ACM | file LICENSE 
##                                             3 
##                                          AGPL 
##                                             2 
##                                        AGPL-3 
##                                            20 
##                         AGPL-3 | file LICENSE 
##                                             2 
##                                   AGPL (>= 3) 
##                                             3 
##                    AGPL (>= 3) + file LICENSE 
##                                             1 
##                        Apache License (== 2.0) 
##                                            16 
##        Apache License (== 2.0) | file LICENSE 
##                                             3
##                     Apache License (>= 2.0)
##                                           3
##                          Apache License 2.0
##                                          16
##           Apache License 2.0 | file LICENSE
##                                           2
##                                Artistic-2.0
##                                          42
##                        Artistic License 2.0
##                                           4
##                                         BSD
##                                          18
##         BSD 2-clause License + file LICENSE
##                                           2
##                        BSD 3-clause License
##                                           1
##         BSD 3-clause License + file LICENSE
##                                           1
##                 BSD_2_clause + file LICENCE
##                                           1
##                 BSD_2_clause + file LICENSE
##                                          35
##    BSD_2_clause + file LICENSE | GPL (>= 2)
##                                           1
##                 BSD_3_clause + file LICENCE
##                                           3
##                 BSD_3_clause + file LICENSE
##                                          59
##         BSD_3_clause + file LICENSE | GPL-2
##                                           1
##                                         BSL
##                                           1
##                                     BSL-1.0
##                                           1
##                          CC BY-NC-ND 3.0 US
##                                           2
##                             CC BY-NC-SA 3.0
##                                           4
##              CC BY-NC-SA 3.0 + file LICENSE
##                                           1
##                             CC BY-NC-SA 4.0
##                                           7
##                 CC BY-NC 3.0 + file LICENSE
##                                           1
##                                CC BY-NC 4.0
##                                           2
##                 CC BY-SA 2.0 + file LICENSE
##                                           1
##                                CC BY-SA 4.0
##                                           3
##                 CC BY-SA 4.0 + file LICENSE
##                                           1
##                                   CC BY 4.0
##                                           1
##                                         CC0
##                                          28
##                                      CeCILL
##                                           6
##                                    CeCILL-2
##                                           7
##                            CeCILL-2 | GPL-2
##                                           1
##           Common Public License Version 1.0
##                                           2
##                                     CPL-1.0
##                                           1
##                                         EPL
##                                           1
##                                EPL (>= 1.0)
##                                           1
##                                        EUPL
##                                           3
##                                file LICENSE
##                                          19
##                                     FreeBSD
##                                           5
##                      FreeBSD | file LICENSE
##                                           1
##              FreeBSD | GPL-2 | file LICENSE
##                                           1
##                  GNU General Public License
##                                          11
##           GNU General Public License (>= 2)
##                                           1
##           GNU General Public License (>= 3)
##                                           1
##                                         GPL
##                                         380
##                                       GPL-2
##                                        1451
##                        GPL-2 | Artistic-2.0
##                                           1
##                        GPL-2 | file LICENCE
##                                           2
##                        GPL-2 | file LICENSE
##                                          14
##                               GPL-2 | GPL-3
##                                         153
## GPL-2 | GPL-3 | BSD_3_clause + file LICENSE
##                                           1
##                  GPL-2 | GPL (>= 2) | GPL-3
##                                           1
##                  GPL-2 | MIT + file LICENCE
##                                           1
##                                       GPL-3
##                                         700
##                        GPL-3 | file LICENSE
##                                          13
##                         GPL-3 + file LICENSE
##                                           19
##                                   GPL (== 2)
##                                            1
##                                    GPL (> 2)
##                                            7
##                                    GPL (> 3)
##                                            3
##                                 GPL (>= 1.0)
##                                            1
##                                   GPL (>= 2)
##                                         2631
##     GPL (>= 2) | BSD_2_clause + file LICENSE
##                                            2
##     GPL (>= 2) | BSD_3_clause + file LICENSE
##                                            1
##                    GPL (>= 2) | file LICENCE
##                                            6
##                    GPL (>= 2) | file LICENSE
##                                           15
##                         GPL (>= 2) | FreeBSD
##                                            1
##      GPL (>= 2) | LGPL (>= 2) | file LICENSE
##                                            1
##                     GPL (>= 2) | LGPL (>= 3)
##                                            1
##                                 GPL (>= 2.0)
##                                           73
##                  GPL (>= 2.0) | file LICENCE
##                                            1
##                                 GPL (>= 2.1)
##                                            1
##                                GPL (>= 2.14)
##                                            2
##                                GPL (>= 2.15)
##                                            1
##                              GPL (>= 2.15.1)
##                                            1
##                                 GPL (>= 2.2)
##                                            1
##                                   GPL (>= 3)
##                                          307
##                    GPL (>= 3) | file LICENCE
##                                            3
##                    GPL (>= 3) | file LICENSE
##                                            5
##                    GPL (>= 3) + file LICENSE
##                                            2
##                                 GPL (>= 3.0)
##                                           15
##                               GPL (>= 3.1.2)
##                                            1
##                GPL | Apache License (== 2.0)
##                                            1
##                          GPL | file LICENSE
##                                           1
##                                        LGPL
##                                          31
##                                      LGPL-2
##                                          12
##             LGPL-2 | LGPL-3 | GPL-2 | GPL-3
##                                           1
##                                    LGPL-2.1
##                                          16
##                     LGPL-2.1 | file LICENSE
##                                           3
##                                      LGPL-3
##                                          72
##                 LGPL-3 | Apache License 2.0
##                                           3
##                       LGPL-3 | file LICENSE
##                                           3
##                       LGPL-3 + file LICENSE
##                                           1
##                                LGPL (> 2.0)
##                                           1
##                                 LGPL (>= 2)
##                                          14
##                  LGPL (>= 2) | file LICENSE
##                                           1
##                               LGPL (>= 2.0)
##                                           7
##                          LGPL (>= 2.0, < 3)
##                                           2
## LGPL (>= 2.0, < 3) | Mozilla Public License
##                                           1
##                               LGPL (>= 2.1)
##                                          27
##                                 LGPL (>= 3)
##                                          14
##                  LGPL (>= 3) | file LICENSE
##                                           1
##                               LGPL (>= 3.0)
##                                           2
##                       Lucent Public License
##                                           1
##                                         MIT
##                                          11
##                          MIT | file LICENSE
##                                           1
##                                 MIT | GPL-2
##                                           1
##                          MIT + file LICENSE
##                                         236
##              MIT + file LICENSE | Unlimited
##                                           1
##                                 MIT License
##                                           2
##                   Mozilla Public License 1.1
##                                            1
##                   Mozilla Public License 2.0
##                                            6
##                                          MPL
##                                            1
##                                      MPL-1.1
##                                            1
##                                 MPL (== 1.1)
##                                            1
##                                 MPL (== 2.0)
##                                            1
##       MPL (>= 2) | GPL (>= 2) | file LICENSE
##                                            1
##                                 MPL (>= 2.0)
##                                            2
##                                    Unlimited
##                                           46
##                     Unlimited | file LICENSE
##                                            1
```

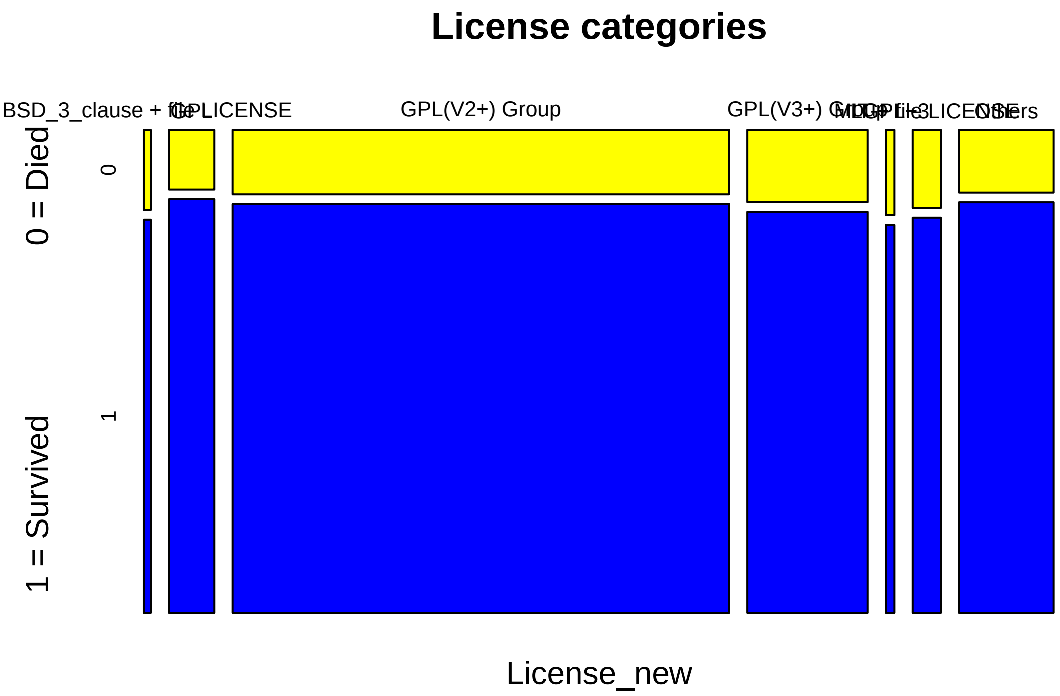
Commentary: In the 2015 data, there are 10 licenses that appear more than 50 times, and 7 of the 10 are GPL-2 or GPL-3 or some combination, which are: `GPL (>= 2)` (with 2631 times), `GPL-2` (with 1451 times), `GPL-3` (with 700 times), `GPL` (with 380 times), `GPL (>= 3)` (with 307 times), `GPL-2 | GPL-3` (with 153 times) and `GPL (>= 2.0)` (with 73 times).

```
License_new <- as.vector(B.df$License)
# Make "GPL-2", "GPL (>= 2)" and "GPL (>= 2.0)" into "GPL (version 2 or later) Group"
Condition1 <- ifelse(B.df$License == "GPL-2"|B.df$License == "GPL (>= 2)"| B.df$License == "GPL (>= 2.0
                     "GPL(V2+) Group", License_new)
# Make "GPL-3" and "GPL (>= 3)" into "GPL (version 3 or later) Group"
Condition2 <- ifelse(B.df$License == "GPL-3"|B.df$License == "GPL (>= 3)", "GPL(V3+) Group", Condition1
# The rest of high frequency categories remains, but make the low frequency categories into "Others"
Condition3 <- ifelse(
Condition2!="BSD_3_clause + file LICENSE"&Condition2!="GPL"&Condition2!="GPL-2|GPL-3"&Condition2!="LGPL-
"Others",
Condition2)
# Create a new variable in the data frame:
B.df$License_new <- Condition3
B.df$License_new <- as.factor(B.df$License_new)
```

```
table(B.df$License_new)
```

```
##
## BSD_3_clause + file LICENSE                         GPL
##                          59                         380
##              GPL(V2+) Group              GPL(V3+) Group
##                        4155                        1007
##                      LGPL-3          MIT + file LICENSE
##                          72                         236
##                      Others
##                         791
```

```
# Mosaic plot of the version Num_Depends v.s. Survived
with(B.df, mosaicplot(table(License_new, Survived), col=c("yellow","blue"),
                      main="License categories", ylab = "1 = Survived                   0 = Died"))
```

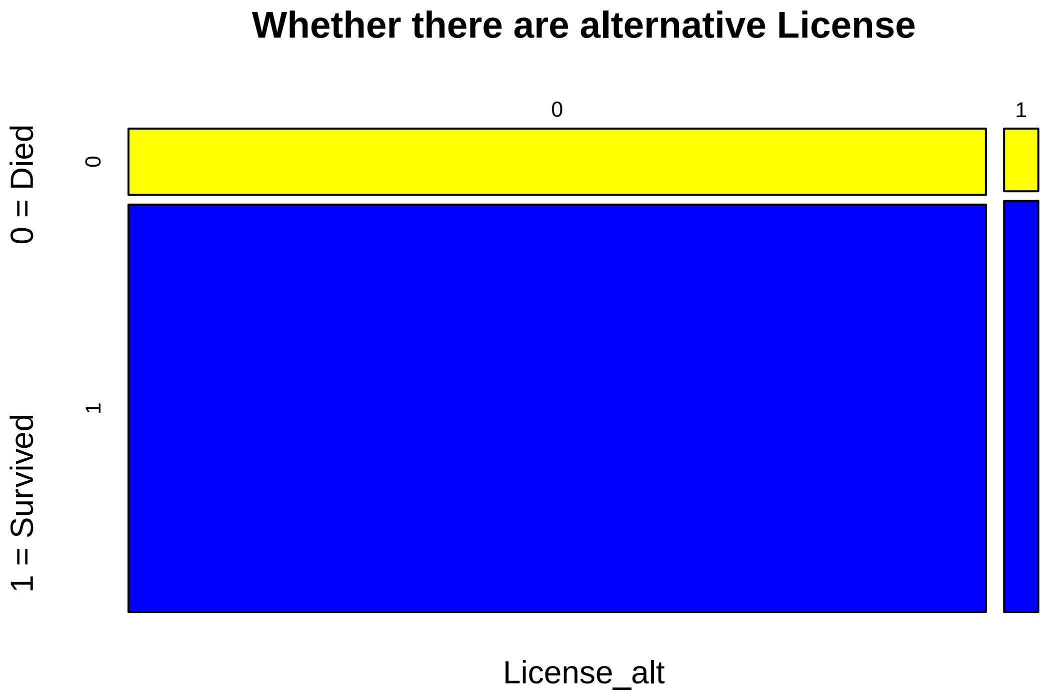


2) Create a binary `License_alt` for whether there are alternatives License: we know the strings will be separated by "|" characters characters if there are alternatives.

```
B.df$License_alt <- as.factor(as.numeric(grepl("[|]", B.df$License)))
```

Visualization:

```
# Mosaic plot of the version License_alt v.s. Survived
with(B.df, mosaicplot(table(License_alt, Survived), col=c("yellow","blue"),
                      main="Whether there are alternative License", ylab = "1 = Survived
```



Commentary: Whether there are alternative license seems affect little on the probability of survival.

Modelling

Implement the binomial regression model:

```r
# Relevel the License_new
B.df <- within(B.df, License_new <- relevel(License_new, ref = "Others"))
```

```r
mod.bin <- glm(Survived~Version_new+Depends_on_version+Num_Depends+License_new+License_alt,
               data = B.df, family = binomial(link = "logit"))
summary(mod.bin)
```

```
##
## Call:
## glm(formula = Survived ~ Version_new + Depends_on_version + Num_Depends +
##     License_new + License_alt, family = binomial(link = "logit"),
##     data = B.df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5272   0.4929   0.5328   0.5870   0.9186
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                               1.82245    0.17305  10.531  < 2e-16 ***
## Version_new1                              0.16623    0.08442   1.969  0.04894 *
## Version_new2                              0.42201    0.15951   2.646  0.00815 **
## Version_new3                              0.19286    0.27839   0.693  0.48846
## Version_new4                              1.18153    0.52314   2.259  0.02391 *
## Version_newothers                         0.62104    0.32587   1.906  0.05667 .
## Depends_on_version1                      -0.04633    0.09501  -0.488  0.62585
## Num_Depends                              -0.13170    0.02336  -5.639 1.71e-08 ***
## License_newBSD_3_clause + file LICENSE -0.26746    0.44090  -0.607  0.54410
## License_newGPL                            0.19391    0.22895   0.847  0.39702
## License_newGPL(V2+) Group                 0.10454    0.14694   0.711  0.47681
## License_newGPL(V3+) Group                -0.08281    0.16918  -0.490  0.62449
## License_newLGPL-3                        -0.16158    0.35033  -0.461  0.64465
## License_newMIT + file LICENSE            -0.05731    0.26292  -0.218  0.82747
## License_alt1                              0.10877    0.24155   0.450  0.65250
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4328.5  on 5184  degrees of freedom
## Residual deviance: 4279.4  on 5170  degrees of freedom
##   (1515 observations deleted due to missingness)
## AIC: 4309.4
##
## Number of Fisher Scoring iterations: 5
```

License doesn't seem to matter. Number of dependencies does, and version might.

```
mod.bin2 <- glm(Survived~Version_new+Depends_on_version+Num_Depends,
                data = B.df, family = binomial(link = "logit"))
summary(mod.bin2)
```

```
##
## Call:
## glm(formula = Survived ~ Version_new + Depends_on_version + Num_Depends,
##     family = binomial(link = "logit"), data = B.df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4712   0.5048   0.5470   0.5805   0.9203
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          1.88176    0.10931  17.215  < 2e-16 ***
## Version_new1         0.17206    0.08412   2.045  0.04081 *
## Version_new2         0.42906    0.15892   2.700  0.00694 **
## Version_new3         0.19977    0.27775   0.719  0.47198
## Version_new4         1.18099    0.52199   2.262  0.02367 *
## Version_newothers    0.65099    0.32509   2.002  0.04523 *
## Depends_on_version1 -0.05779    0.09443  -0.612  0.54059
## Num_Depends         -0.12853    0.02318  -5.545 2.94e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

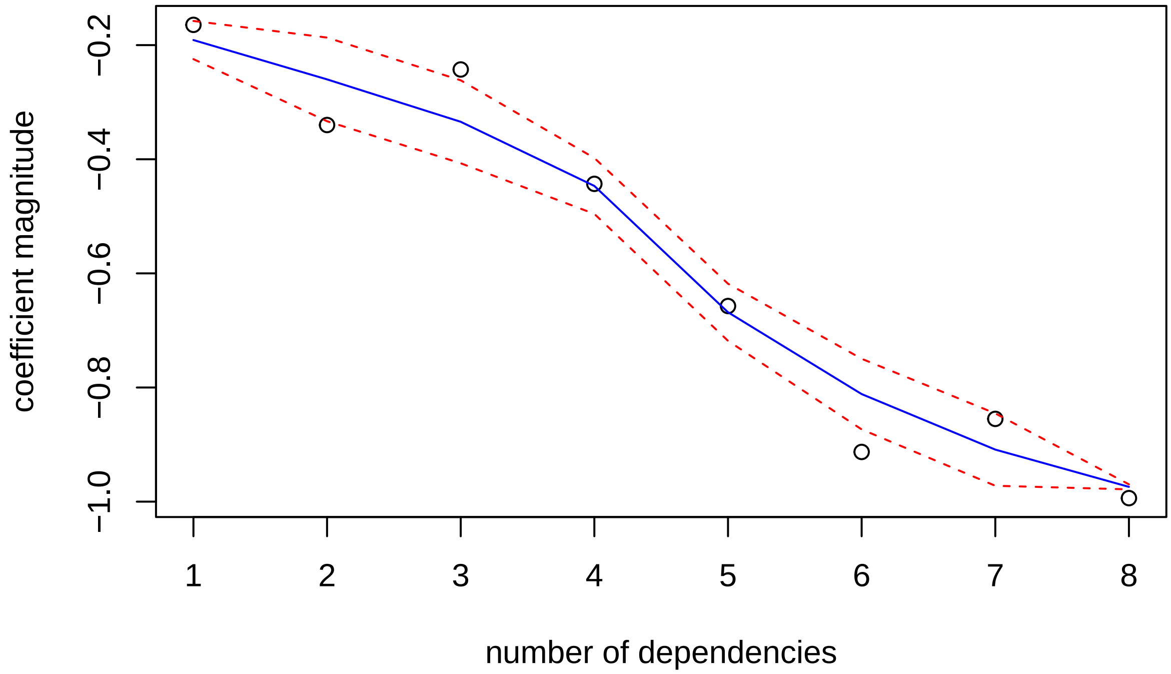
```
##     Null deviance: 4328.5  on 5184  degrees of freedom
## Residual deviance: 4284.3  on 5177  degrees of freedom
##   (1515 observations deleted due to missingness)
## AIC: 4300.3
##
## Number of Fisher Scoring iterations: 5
```

It seems that versioned dependencies doesn't matter much.

```
mod.bin3 <- glm(Survived~factor(Num_Depends)+Version_new,
                data = B.df, family = binomial(link = "logit"))
summary(mod.bin3)
```

```
##
## Call:
## glm(formula = Survived ~ factor(Num_Depends) + Version_new, family = binomial(link = "logit"),
##     data = B.df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4629   0.4992   0.5414   0.5847   1.1774
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)             1.84616    0.08862  20.833  < 2e-16 ***
## factor(Num_Depends)1   -0.16450    0.10924  -1.506 0.132103
## factor(Num_Depends)2   -0.33999    0.11788  -2.884 0.003925 **
## factor(Num_Depends)3   -0.24269    0.14438  -1.681 0.092788 .
## factor(Num_Depends)4   -0.44310    0.15331  -2.890 0.003850 **
## factor(Num_Depends)5   -0.65720    0.19866  -3.308 0.000939 ***
## factor(Num_Depends)6   -0.91293    0.27720  -3.293 0.000990 ***
## factor(Num_Depends)7   -0.85499    0.52323  -1.634 0.102245
## factor(Num_Depends)8   -0.99383    0.52920  -1.878 0.060382 .
## factor(Num_Depends)9    9.58249  324.74412   0.030 0.976460
## factor(Num_Depends)10 -14.41222  229.62849  -0.063 0.949955
## factor(Num_Depends)11  -2.01968    1.41710  -1.425 0.154092
## Version_new1            0.17353    0.08424   2.060 0.039416 *
## Version_new2            0.43022    0.15915   2.703 0.006865 **
## Version_new3            0.19770    0.27805   0.711 0.477063
## Version_new4            1.13741    0.52220   2.178 0.029397 *
## Version_newothers       0.64148    0.32525   1.972 0.048579 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4328.5  on 5184  degrees of freedom
## Residual deviance: 4277.1  on 5168  degrees of freedom
##   (1515 observations deleted due to missingness)
## AIC: 4311.1
##
## Number of Fisher Scoring iterations: 11
```

```
trendscatter(1:8, coef(mod.bin3)[2:9], xlab = "number of dependencies",
             ylab = "coefficient magnitude", main = "")
```



It looks like a linear decrease in general.

```
B.df$Version_new0<-ifelse(B.df$Version_new %in% c(1, 2, 3, 4), "1-4", B.df$Version_new)
mod.bin4 <- glm(Survived~Num_Depends+Version_new0, data = B.df, family = binomial(link = "logit"))
summary(mod.bin4)
```

```
##
## Call:
## glm(formula = Survived ~ Num_Depends + Version_new0, family = binomial(link = "logit"),
##     data = B.df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2580   0.4915   0.5465   0.5785   0.8973
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.82581    0.07255  25.168  < 2e-16 ***
## Num_Depends     -0.12279    0.02287  -5.368 7.94e-08 ***
## Version_new01-4  0.22675    0.08034   2.822  0.00477 **
## Version_new06    0.64209    0.32479   1.977  0.04805 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4328.5  on 5184  degrees of freedom
## Residual deviance: 4291.9  on 5181  degrees of freedom
##   (1515 observations deleted due to missingness)
## AIC: 4299.9
##
## Number of Fisher Scoring iterations: 4
```

The exponentiated estimates for variables:

```
round(exp(coef(mod.bin4)), 3)
```

```
##     (Intercept)     Num_Depends Version_new01-4   Version_new06
##           6.208           0.884           1.255           1.900
```

The exponentiated confidence intervals for variables:

```
round(exp(confint(mod.bin4)), 3)
```

```
## Waiting for profiling to be done...
```

```
##                 2.5 % 97.5 %
## (Intercept)     5.394  7.168
## Num_Depends     0.846  0.925
## Version_new01-4 1.071  1.468
## Version_new06   1.051  3.800
```

Summary

License doesn't seem to be related to package survival. Packages with more dependences were less likely to survive, by a factor of 0.88 per dependency; versioned dependencies didn't matter much.

Packages in version 0 were the least likely to have survived. Those with numeric versions starting 1, 2, 3, and 4 were about 1.26 as likely, and those with other numbering schemes were substantially more likely (probably because these represented people with pre-existing versioning policies from experience with programming).

Task Three

Describe how dependencies, and license relate to the subsequent survival time of a package on CRAN in 2015

Data Cleaning

```
# Rename the column in B.df to make it consistent with all_dates'
colnames(B.df)[1] <- "pkg"
```

```
# Combine the 2 data frame into one:
combined.df <- merge(all_dates, B.df, all=TRUE)
# Delete those records 'Survived' information is NA:
combined.df <- subset(combined.df, !is.na(Survived))
```

Modelling

Implement the Cox model:

```
modelcox1 <- coxph(Surv(end_date - as.numeric(as.Date("2015-06-02")), Survived==0)~Version_new+Depends_
modelcox1
```

```
## Call:
## coxph(formula = Surv(end_date - as.numeric(as.Date("2015-06-02")),
##     Survived == 0) ~ Version_new + Depends_on_version + Num_Depends +
##     License_new + License_alt, data = combined.df)
##
##                                            coef exp(coef) se(coef)      z
## Version_new1                           -0.14449   0.86547  0.07743 -1.866
## Version_new2                           -0.39177   0.67586  0.14875 -2.634
## Version_new3                           -0.17755   0.83732  0.25656 -0.692
## Version_new4                           -1.12001   0.32628  0.50458 -2.220
## Version_newothers                      -0.56714   0.56714  0.30762 -1.844
## Depends_on_version1                     0.02651   1.02686  0.08738  0.303
## Num_Depends                             0.12275   1.13060  0.02109  5.821
## License_newBSD_3_clause + file LICENSE  0.16927   1.18443  0.39862  0.425
## License_newGPL                         -0.21066   0.81005  0.21188 -0.994
## License_newGPL(V2+) Group              -0.11482   0.89153  0.13469 -0.852
## License_newGPL(V3+) Group               0.04087   1.04172  0.15463  0.264
## License_newLGPL-3                       0.02178   1.02202  0.32780  0.066
## License_newMIT + file LICENSE           0.02111   1.02133  0.24111  0.088
## License_alt1                           -0.13097   0.87724  0.22245 -0.589
##                                               p
## Version_new1                            0.06204
## Version_new2                            0.00845
## Version_new3                            0.48892
## Version_new4                            0.02644
## Version_newothers                       0.06523
## Depends_on_version1                     0.76163
## Num_Depends                            5.84e-09
## License_newBSD_3_clause + file LICENSE  0.67111
## License_newGPL                          0.32011
## License_newGPL(V2+) Group               0.39395
## License_newGPL(V3+) Group               0.79152
## License_newLGPL-3                       0.94702
## License_newMIT + file LICENSE           0.93024
## License_alt1                            0.55603
##
## Likelihood ratio test=49.46  on 14 df, p=7.503e-06
## n= 5184, number of events= 761
##    (1516 observations deleted due to missingness)
```

```
modelcox2 <-coxph(Surv(end_date - as.numeric(as.Date("2015-06-02")), Survived==0)~Version_new0+Num_Depe
modelcox2
```

```
## Call:
## coxph(formula = Surv(end_date - as.numeric(as.Date("2015-06-02")),
##     Survived == 0) ~ Version_new0 + Num_Depends, data = combined.df)
##
##                     coef exp(coef) se(coef)      z        p
## Version_new01-4 -0.20069   0.81817  0.07377 -2.720  0.00652
## Version_new06   -0.58608   0.55650  0.30679 -1.910  0.05608
## Num_Depends      0.11525   1.12216  0.02067  5.577 2.45e-08
##
## Likelihood ratio test=37.4  on 3 df, p=3.785e-08
## n= 5184, number of events= 761
##    (1516 observations deleted due to missingness)
```

Summary

The hazard ratio for `Num_Depends` is 1.122, so packages with a 1 number higher in the dependency on other packages had a 12.2% higher rate (hazard) of death (on average, over the follow-up time).

The hazard ratio for a 2015 package in version category '1-4' to version category '0' (baseline group) is 0.818, so package in version category '1' had a 18.2% lower rate (hazard) of death than the baseline group (on average, over the follow-up time);

The hazard ratio for a 2015 package in version category 'others' to version category '0' (baseline group) is 0.557, so package in version category '1' had a 44.3% lower rate (hazard) of death than the baseline group (on average, over the follow-up time);

In conclusion, the Cox model gives a very similar message to the logistic model. The coefficients have similar magnitudes (but opposite signs) and so do the standard errors.