School of Statistics, University of the Philippines (Diliman)
Linangan ng Estadistika, Unibersidad ng Pilipinas (Diliman)

25 of November 2015

Al-Ahmadgaid B. Asaad

Problem Set 6 | **Stat 245**

7.9 In Table D.3 of Appendix D, data is reported on the death times of 863 kidney transplant patients (see section 1.7). Here, patients can be classified by race and sex into one of the four groups.

(a) Test the hypothesis that there is no difference in survival between the four groups.

(b) Adjusting for the sex of the patients, test the hypothesis that blacks have a higher mortality rate than whites. Also provide individual tests, for each sex, of the hypothesis of no racial differences in survival rates.

Solution

(a) There are four groups in this problem, 432 white males, 92 black males, 280 white females, and 59 black females, in total there are 863 patients. Here we want to test the null hypothesis that the hazard rates between these group are equal for all $t \leq \tau$ versus the alternative that at least one of the hazard rates is different for some $t \leq \tau$. In order to run the test in R, we need the data in section 1.7 named as `kidtran` from `KMsurv` package, thus we need to install `KMsurv`:

```
> install.packages("KMsurv")
```

Once successfully installed, load the package and the data

```
> library(KMsurv)
> data(kidtran)
```

To view the first 6 observations use the `head` function:

```
> head(kidtran)
  obs time delta gender race age
1   1    1     0      1    1  46
2   2    5     0      1    1  51
3   3    7     1      1    1  55
4   4    9     0      1    1  57
5   5   13     0      1    1  45
6   6   13     0      1    1  43
```

The following are descriptions of the columns of the data above, (access this using `help(kidtran)` or `?kidtran`):

Columns	Descriptions
`obs`	Observation number
`time`	Time to death or on-study time
`delta`	Death indicator (0 = alive, 1 = dead)
`gender`	1 = male, 2 = female
`race`	1 = white, 2 = black
`age`	Age in years

Using survival R package

Now for testing, we'll consider first a built-in package in R called `survival`. In this package, we're gonna use `survdiff` and `Surv` functions (run `?survdiff` and `?Surv` to get help on these functions), so that the following is the code for testing the hypotheses

```
> survdiff(Surv(time, delta) ~ race + gender,
+          data = kidtran, rho = 0)
```

and below is the output

```
Call:
survdiff(formula = Surv(time, delta) ~ race + gender,
         data = kidtran, rho = 0)

                      N Observed Expected (O-E)^2/E (O-E)^2/V
race=1, gender=1 432       73    69.25    0.2025    0.4013
race=1, gender=2 280       39    47.39    1.4860    2.2531
race=2, gender=1  92       14    14.52    0.0184    0.0205
race=2, gender=2  59       14     8.84    3.0173    3.2245

 Chisq= 4.7  on 3 degrees of freedom, p= 0.192
```

The `Observed` column above is obtained by $\sum_i d_{ij}$, and the `Expected` column is obtained by $\sum_i Y_{ij}\left(\frac{d_i}{Y_i}\right)$, and then supply this to the third column. Unfortunately, I don't know how to compute the fourth column and that's the disadvantage of using an R package, you don't know what's going on behind the scene. Anyway, the test statistic is $\chi^2_3 = 4.7$, with p-value greater than .05 (say this is our level of significance). Therefore, there is no significant difference between the hazard rates of the four groups. And by the way, the weight function above is $\hat{S}(t_{i-1})^{\text{rho}}$ and since `rho = 0` above, then it's just a log-rank weight test.

Using custom R function

An alternative to above code is to program the theory detailed in Section 7.3 so that we can explore Z_j and the variance-covariance matrix $\mathbf{\Sigma}$. In the R script attached with this pdf, the function for computing Z_j is `z_j` with the following usage,

```
z_j(x, K, di1_col, Yi1_col)
```

where `x` is the pivot data; `K` is the group size; `di1_col` is the column of the first d_i (observed event of first group); `Yi1_col` is the column of the first Y_i (individual at risk of first group). And for variance-covariance matrix $\mathbf{\Sigma}$ is `sigma` with the following usage,

```
sigma(x, K, Yi1_col)
```

the descriptions of the parameters above is similar to that in `z_j` function. Now in order to obtain the pivot data `x` and to determine `di1_col` and `Yi1_col`, a preliminary data preparation function is also included in the R script, the function is named as `data_setup` with the following usage,

```
data_setup(x, time, covariates, grp_names = NULL)
```

where `x` is the raw data; `time` is the time variable in the raw data; `covariates` is the covariates in the raw data, this is a character vector; `grp_names` is the group names that depends on the arrangement of the `covariates`.

Let's perform the test, using the `kidtran` data, for four groups (`K = 4`), with time as `time` variable of `kidtran`, covariates as `race` and `gender`, and group names (`grp_names`) as `wm` (white(1) males(1)), `wf` (white(1) females(2)), `bm` (black(2) males(1)), and `bf` (black(2) females(2)). Notice the group names depend on the arrangement of the covariates, in this case `race` first before `gender`, so if the first group is (males(1) white(1)) then that's misleading. Setup the pivot data as follows,

```
> covar <- c('delta', 'race', 'gender')
> g_names = c('wm', 'wf', 'bm', 'bf')
> x_data <- data_setup(x = kidtran, time = 'time',
+                      covariates = covar,
+                      grp_names = g_names)
> head(x_data)
  time ci_wm ci_wf ci_bm ci_bf di_wm di_wf
1    1     1     1     0     0     0     0
2    2     0     0     0     0     0     1
3    3     0     0     0     0     0     1
4    5     1     1     0     0     0     0
5    7     0     0     0     0     1     1
6    9     1     1     0     0     0     0
  di_bm di_bf Yi_wm Yi_wf Yi_bm Yi_bf  Yi di
1     0     0   432   280    92    59 863  0
2     0     0   431   279    92    59 861  1
3     0     0   431   278    92    59 860  1
4     0     0   431   277    92    59 859  0
5     0     0   430   276    92    59 857  2
6     0     0   429   275    92    59 855  0
```

So the output is a data frame consisting of 15 columns, the four `ci`'s column for censored observations in the four groups, the `di_wm`, `di_wf`, `di_bm`, `di_bf`, columns for d_{ij}, the `Yi_wm`, `Yi_wf`, `Yi_bm` `Yi_bf` for Y_{ij}, the `Yi` column for Y_i, and the `di` column for d_i.

To obtain the Z_j, we'll use the `x_data`, and supply it to the following code:

```
> zj <- z_j(x = x_data, K = 4, di1_col = 6, Yi1_col = 10)
> zj
           [,1]
[1,]  3.7450042
[2,] -8.3918316
[3,] -0.5167172
[4,]  5.1635446
```

`K = 4` since there are four groups, and `di1_col = 6` since the `di_wm` is in the sixth column of `x_data`, `Yi1_col = 10` since the `Yi_wm` is in the tenth column of `x_data`. Next is to compute the variance-covariance matrix, $\mathbf{\Sigma}$ as follows

```
> Sigma <- sigma(x = x_data, K = 4, Yi1_col = 10)
> Sigma
           [,1]       [,2]       [,3]       [,4]
[1,]  34.949762 -23.387739 -7.1833754 -4.3786478
[2,] -23.387739  31.256016 -4.8985901 -2.9696867
[3,]  -7.183375  -4.898590 13.0023724 -0.9204069
[4,]  -4.378648  -2.969687 -0.9204069  8.2687415
```

Therefore the test statistics given by

$$\chi^2 = [Z_1(\tau), \cdots, Z_{K-1}(\tau)]\mathbf{\Sigma}^{-1}[Z_1(\tau), \cdots, Z_{K-1}(\tau)]^T$$

is coded as,

```
> t(zj[1:3, ]) %*% solve(Sigma[1:3, 1:3]) %*% zj[1:3, ]
        [,1]
[1,] 4.73631
```

and the p-value is,

```
> 1 - pchisq(4.73631, 3)
[1] 0.1921559
```

To wrap-up the process above, the last function in the R script will do just that. The function is named as `survTest`, so that

```
> survTest(x = x_data, K = 4, di1_col = 6, Yi1_col = 10)
$`Chi-square`
        [,1]
[1,] 4.73631

$`p-value`
          [,1]
[1,] 0.1921559
```

and we obtain the same result for test statistics and p-value as that in the `survdiff` function of the `survival` package.

(b) In this problem, we need to stratify the data (`kidtran`) with respect to `gender`, and also do the test within each stratum. Let's do the latter part first. To filter the data with respect to gender consider the following code,

```
> males <- kidtran[kidtran[, 'gender'] == 1, ]
```

So our new data is named as `males`, and therefore we are testing between two samples (white males, and black males).

Using survival R package

The following will test the difference between the hazard rates of the two samples,

```
> survdiff(Surv(time, delta) ~ race, data = males)
Call:
survdiff(formula = Surv(time, delta) ~ race, data = males)

         N Observed Expected (O-E)^2/E (O-E)^2/V
race=1 432       73     71.9    0.0168     0.097
race=2  92       14     15.1    0.0801     0.097

 Chisq= 0.1  on 1 degrees of freedom, p= 0.755
```

So the test statistic $\chi^2_1 = .1$ with p-value .755, which is greater than .05 (say this is our level of significance), then it simply suggests that there is no difference between the hazard rates of the two groups (white males, and black males). Using the same approach for female stratum (white females, and black female), we have the following output

```
> females <- kidtran[kidtran[, 'gender'] == 2, ]
> survdiff(Surv(time, delta) ~ race, data = females)
Call:
survdiff(formula = Surv(time, delta) ~ race, data = females)

         N Observed Expected (O-E)^2/E (O-E)^2/V
race=1 280       39    44.79     0.748      4.85
race=2  59       14     8.21     4.076      4.85

 Chisq= 4.8  on 1 degrees of freedom, p= 0.0277
```

The p-value is .0277 less than .05 so there is a significant difference between the hazard rates (for white females and black female).

Now to test the hypothesis that blacks have a higher mortality rate than whites, we use the `strata` function and apply this to the `gender` variable, that is

```
> survdiff(Surv(time, delta) ~ race + strata(gender),
+          data = kidtran)
Call:
survdiff(formula = Surv(time, delta) ~ race + strata(gender),
    data = kidtran)

         N Observed Expected (O-E)^2/E (O-E)^2/V
```

```
race=1 712      112    116.7      0.188       1.13
race=2 151       28     23.3      0.942       1.13

 Chisq= 1.1  on 1 degrees of freedom, p= 0.287
```

and it indicates that there is no enough evidence that blacks have higher mortality rate than whites since the p-value of .287 is greater than .05.

Using custom R function

We can also achieved above computations using the custom functions we programmed, for males

```
> covar <- c(’delta’, ’race’)
> g_names = c(’wm’, ’bm’)
> x_males <- data_setup(x = males, time = ’time’,
+                         covariates = covar,
+                         grp_names = g_names)
> survTest(x = x_males, K = 2, di1_col = 4, Yi1_col = 6)
$‘Chi-square‘
           [,1]
[1,] 0.09702603

$‘p-value‘
          [,1]
[1,] 0.7554281
```

and for females,

```
> covar <- c(’delta’, ’race’)
> g_names = c(’wf’, ’bf’)
> x_females <- data_setup(x = females, time = ’time’,
+                           covariates = covar,
+                           grp_names = g_names)
> survTest(x = x_females, K = 2, di1_col = 4, Yi1_col = 6)
$‘Chi-square‘
         [,1]
[1,] 4.847488

$‘p-value‘
           [,1]
[1,] 0.02768642
```

Now for stratified test, according to the theory we simply solve for Z_{js} and $\mathbf{\Sigma}_s$ for $s = \{\text{males, females}\}$, and then compute the pooled $Z_j = \sum_s Z_{js}$ and pooled $\mathbf{\Sigma} = \sum_s \mathbf{\Sigma}_s$. So let’s compute for Z_{js} as follows:

```
> (z_males <- z_j(x = x_males, K = 2,
+                 di1_col = 4, Yi1_col = 6))
          [,1]
[1,]  1.099833
[2,] -1.099833
> (z_females <- z_j(x = x_females, K = 2,
+                   di1_col = 4, Yi1_col = 6))
          [,1]
[1,] -5.786174
[2,]  5.786174
```

Next we compute for $\mathbf{\Sigma}_s$

```
> (Sigma_males <- sigma(x = x_males, K = 2, Yi1_col = 6))
          [,1]      [,2]
[1,]  12.46708 -12.46708
[2,] -12.46708  12.46708
> (Sigma_females <- sigma(x = x_females, K = 2, Yi1_col = 6))
         [,1]     [,2]
[1,]  6.90663 -6.90663
[2,] -6.90663  6.90663
```

So that the pooled Z_j and $\mathbf{\Sigma}$ is

```
> (zj <- z_males + z_females)
          [,1]
[1,] -4.686341
[2,]  4.686341
> (Sigma <- Sigma_males + Sigma_females)
          [,1]      [,2]
[1,]  19.37371 -19.37371
[2,] -19.37371  19.37371
```

And therefore the χ^2 test statistics for stratified test in this case for two-sample is given by

$$\chi^2 = \frac{\sum_{s=1}^{M} Z_1(\tau)}{\sqrt{\sum_{s=1}^{M} \hat{\sigma}_{11s}}} = \sqrt{[Z_1(\tau)]\sigma_{11}^{-1}[Z_1(\tau)]^T}$$

which in R is equivalent to

```
> (chi <- sqrt(t(zj[1,]) %*% solve(Sigma[1,1]) %*% zj[1,]))
       [,1]
[1,] 1.0647
```

with p-value

```
> 1 - pchisq(chi, 1)
          [,1]
[1,] 0.3021456
```

The output is consistent with that of `survdiff` result using `strata` function, but this one is more precise (with respect to the decimal points).

The codes are all available in the R script.