Survival Analysis with R: Exercises

Exercise set 1

Take a look at the built in `colon` dataset. If you type `?colon` it'll ask you if you wanted help on the colon dataset from the survival package, or the colon operator. Click "Chemotherapy for Stage B/C colon cancer", or be specific with `?survival::colon`. This dataset has survival and recurrence information on 929 people from a clinical trial on colon cancer chemotherapy. There are two rows per person, indidicated by the event type (`etype`) variable – `etype==1` indicates that row corresponds to recurrence; `etype==2` indicates death.

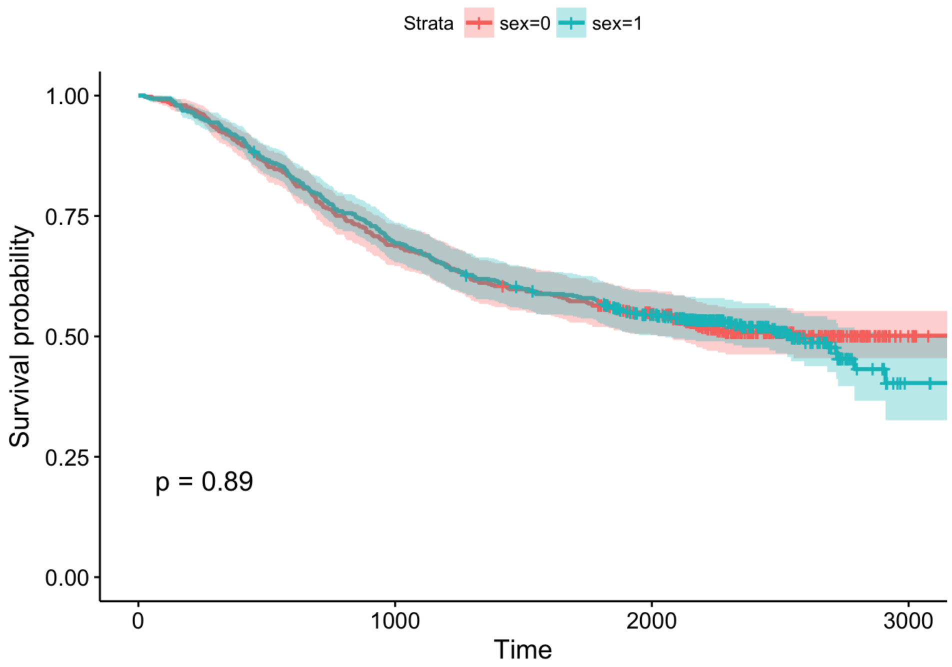
First, let's turn the colon data into a tibble, then filter the data to only include the survival data, not the recurrence data. Let's call this new object `colondeath`. The `filter()` function is in the **dplyr** library, which you can get by running `library(dplyr)`. If you don't have dplyr you can use the base `subset()` function instead.

```r
library(dplyr)
colon <- as_tibble(colon)
colondeath <- filter(colon, etype==2)

# Or, using base subset()
# colondeath <- subset(colon, etype==2)

head(colondeath)
```

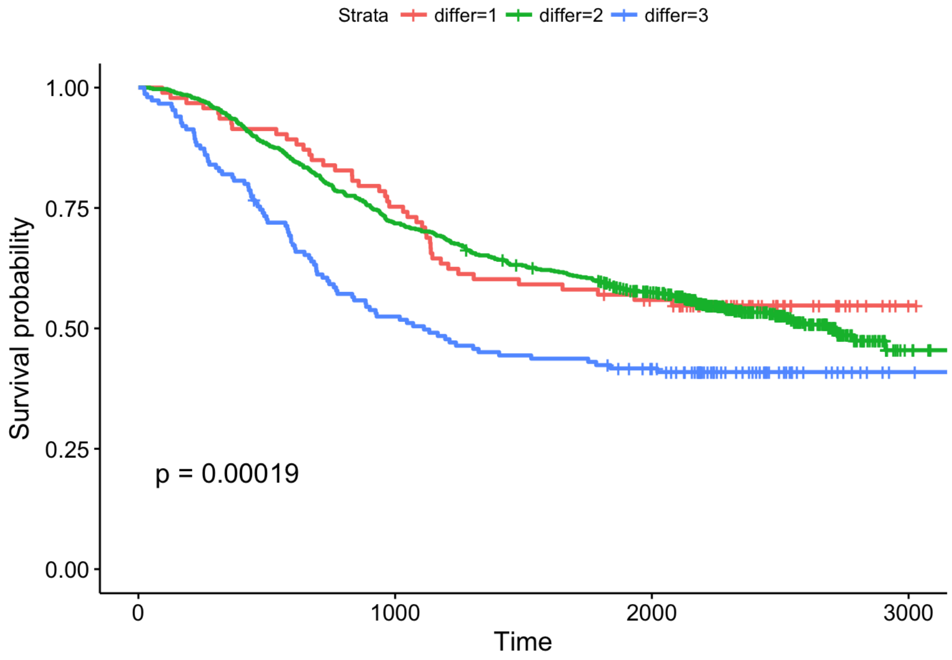
1. Look at the help for `?colon` again. How are `sex` and `status` coded? How is this different from the lung data?
2. Using `survfit(Surv(..., ...,)~..., data=colondeath)`, create a survival curve separately for males versus females. Call the resulting object `sfit`. Run a `summary()` on this object, showing time points 0, 500, 1000, 1500, and 2000. Do males or females appear to fair better over this time period?

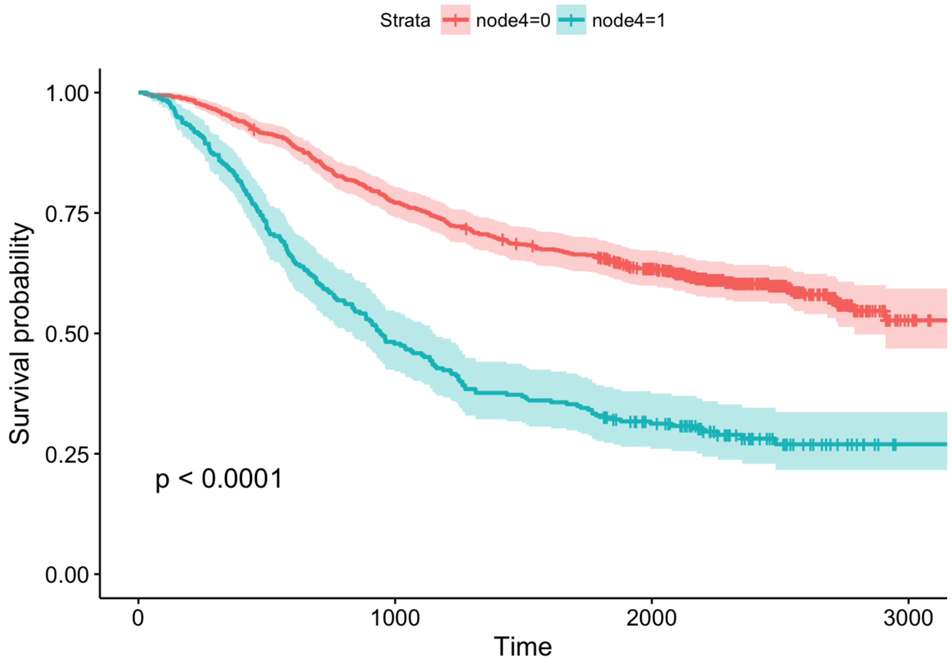
```
##                 sex=0
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     0    445       0    1.000  0.0000        1.000        1.000
##   500    381      64    0.856  0.0166        0.824        0.889
##  1000    306      75    0.688  0.0220        0.646        0.732
##  1500    265      40    0.598  0.0232        0.554        0.645
##  2000    218      22    0.547  0.0236        0.503        0.596
##
##                 sex=1
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     0    484       0    1.000  0.0000        1.000        1.000
##   500    418      65    0.866  0.0155        0.836        0.897
##  1000    335      83    0.694  0.0210        0.654        0.736
##  1500    287      46    0.598  0.0223        0.556        0.644
##  2000    238      25    0.545  0.0227        0.503        0.592
```

3. Using the survminer package, plot a Kaplan-Meier curve for this analysis with confidence intervals and showing the p-value. See `?ggsurvplot` for help. Is there a significant difference between males and females?



4. Create Kaplan-Meier plot stratifying by:
    a. The extent of differentiation (well, moderate, poor), showing the p-value.
    b. Whether or not there was detectable cancer in >=4 lymph nodes, showing the p-value and confidence bands.





Exercise set 2

Let's go back to the `colon` cancer dataset. Remember, you created a `colondeath` object in the first exercise that only includes survival (`etype==2`), not recurrence data points. See `?colon` for more information about this dataset.

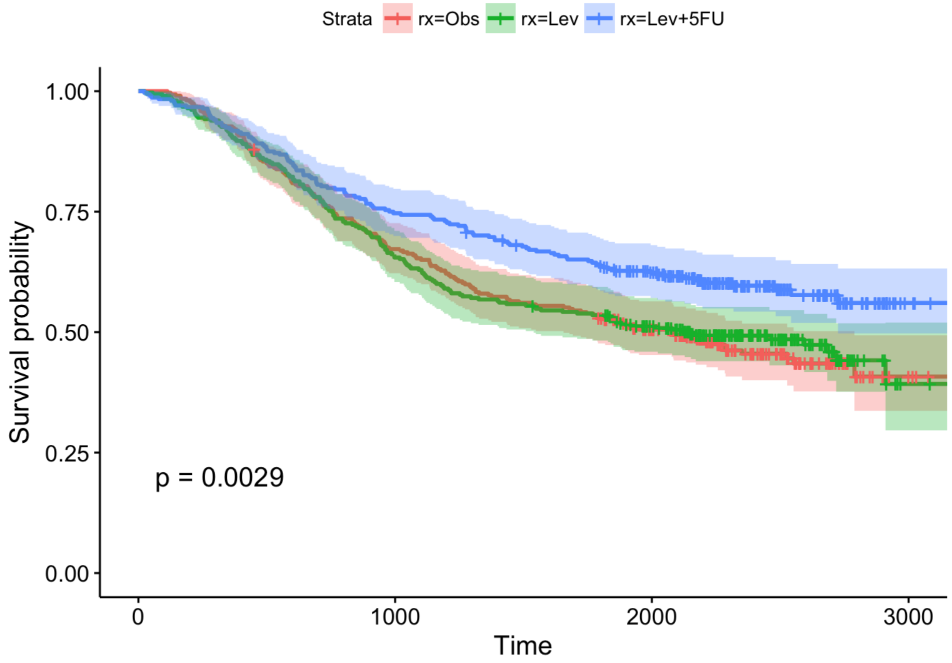
1. Take a look at `levels(colondeath$rx)`. This tells you that the `rx` variable is the type of treatment the patient was on, which is either nothing (coded `Obs`, short for Observation), Levamisole (coded `Lev`), or Levamisole + 5-fluorouracil (coded `Lev+5FU`). This is a factor variable coded with these levels, in that order. This means that `Obs` is treated as the baseline group, and other groups are dummy-coded to represent the respective group.

rx	Lev	Lev+5FU
Obs	0	0
Lev	1	0
Lev+5FU	0	1

2. Run a Cox proportional hazards regression model against this `rx` variable. How do you interpret the result? Which treatment seems to be significantly different from the control (`Obs`ervation)?

```
##                 coef exp(coef) se(coef)     z      p
## rxLev     -0.0266    0.9737   0.1103 -0.24 0.8092
## rxLev+5FU -0.3717    0.6896   0.1188 -3.13 0.0017
##
## Likelihood ratio test=12.2  on 2 df, p=0.0023
## n= 929, number of events= 452
```

3. Show the results using a Kaplan-Meier plot, with confidence intervals and the p-value.



4. Fit another Cox regression model accounting for age, sex, and the number of nodes with detectable cancer. Notice the test statistic on the likelihood ratio test becomes much larger, and the overall model becomes more significant. What do you think accounted for this increase in our ability to model survival?

```
##                coef exp(coef) se(coef)     z       p
## rxLev      -0.08007   0.92305  0.11161 -0.72 0.47312
## rxLev+5FU -0.40253   0.66863  0.12054 -3.34 0.00084
## age         0.00533   1.00535  0.00405  1.32 0.18739
## sex        -0.02826   0.97214  0.09573 -0.30 0.76786
## nodes       0.09275   1.09719  0.00887 10.46 < 2e-16
##
## Likelihood ratio test=87.8  on 5 df, p=0
## n= 911, number of events= 441
##    (18 observations deleted due to missingness)
```

Exercise set 3

The “KIPAN” cohort (in KIPAN.clinical) is the pan-kidney cohort, consisting of KICH (chromaphobe renal cell carcinoma), KIRC (renal clear cell carcinoma), and KIPR (papillary cell carcinoma). The KIPAN.clinical has KICH.clinical, KIRC.clinical, and KIPR.clinical all combined.

1. Using survivalTCGA(), create a new object called clinkid using the KIPAN.clinical cohort. For the columns to extract, get both the disease code and the patient’s gender (extract.cols=c("admin.disease_code", "patient.gender")). The first few rows will look like this.

```
##   times bcr_patient_barcode patient.vital_status admin.disease_code
## 1  1158        TCGA-KL-8323                    1               kich
## 2  4311        TCGA-KL-8324                    0               kich
## 3   725        TCGA-KL-8325                    1               kich
## 4  3322        TCGA-KL-8326                    0               kich
## 5  3553        TCGA-KL-8327                    0               kich
## 6  3127        TCGA-KL-8328                    0               kich
##   patient.gender
## 1         female
## 2         female
## 3         female
## 4           male
## 5         female
## 6           male
```

2. The `xtabs()` command will produce tables of counts for categorical variables. Here's an example for how to use `xtabs()` for the built-in colon cancer dataset, which will tell you the number of samples split by sex and by treatment.

```
xtabs(~rx+sex, data=colon)
```

```
##          sex
## rx          0   1
##   Obs     298 332
##   Lev     266 354
##   Lev+5FU 326 282
```

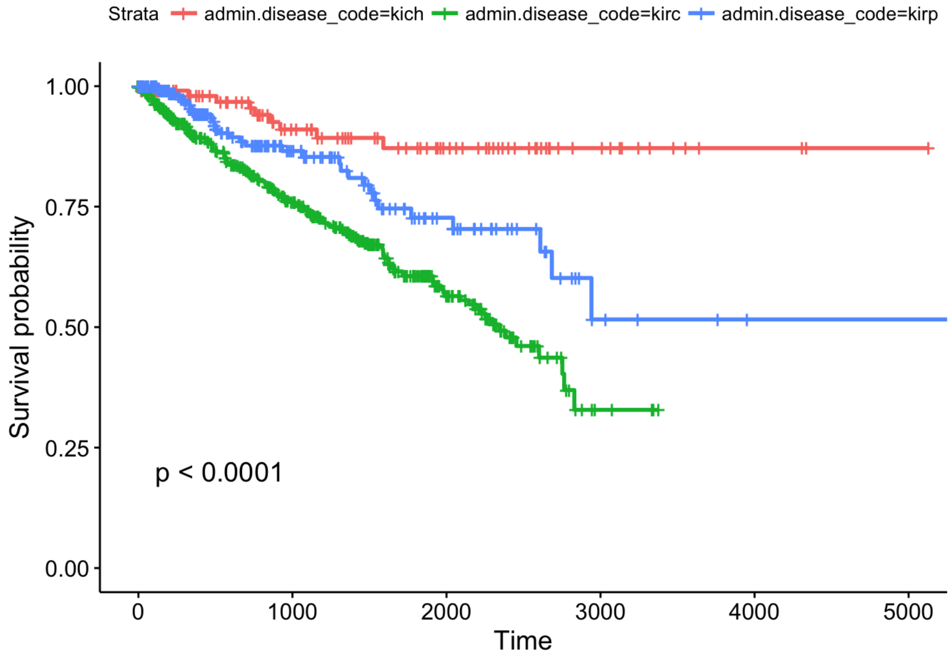
Use the same command to examine how many samples you have for each kidney sample type, separately by sex.

```
##                   patient.gender
## admin.disease_code female male
##               kich     51   61
##               kirc    191  346
##               kirp     76  212
```

3. Run a Cox PH regression on the cancer type and gender. What's the effect of gender? Is it significant? How does survival differ by each type? Which has the worst prognosis?

```
##                            coef exp(coef) se(coef)     z       p
## admin.disease_codekirc  1.5929    4.9179   0.3450  4.62 3.9e-06
## admin.disease_codekirp  0.9962    2.7080   0.3807  2.62  0.0089
## patient.gendermale     -0.0628    0.9391   0.1484 -0.42  0.6721
##
## Likelihood ratio test=39.4  on 3 df, p=1.4e-08
## n= 937, number of events= 203
```

4. Create survival curves for each different subtype.
   a. Produce a Kaplan-Meier plot.
   b. Show survival tables each year for the first 5 years.



```
## Call: survfit(formula = Surv(times, patient.vital_status) ~ admin.disease_code,
##     data = clinkid)
##
##                 admin.disease_code=kich
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     0    111       0    1.000  0.0000        1.000        1.000
##   365     86       2    0.980  0.0144        0.952        1.000
##   730     72       2    0.954  0.0226        0.911        0.999
##  1095     54       3    0.910  0.0329        0.848        0.977
##  1460     44       1    0.893  0.0366        0.824        0.967
##  1825     38       1    0.871  0.0415        0.794        0.957
##
##                 admin.disease_code=kirc
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     0    536       0    1.000  0.0000        1.000        1.000
##   365    385      49    0.895  0.0142        0.868        0.924
##   730    313      32    0.816  0.0186        0.781        0.853
##  1095    250      26    0.744  0.0217        0.703        0.788
##  1460    181      20    0.678  0.0243        0.633        0.728
##  1825    112      16    0.606  0.0277        0.554        0.663
##
##                 admin.disease_code=kirp
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     0    288       0    1.000  0.0000        1.000        1.000
##   365    145      10    0.941  0.0182        0.906        0.977
##   730    100       8    0.877  0.0278        0.824        0.933
##  1095     67       2    0.853  0.0316        0.793        0.917
##  1460     54       3    0.810  0.0388        0.737        0.889
##  1825     36       5    0.727  0.0495        0.636        0.831
```