BMI713 Problem Set 5

Instructions:

Please submit this problem set before class on Tuesday, December 5. Problem sets may be submitted within a week past the due date at a 20% penalty; each person is allowed to submit one problem late (within a week) without penalty. Please comment your code indicating what your functions do and any relevant passage (not necessarily every line of code), because it is part of the requirements of each exercise. Missing comments will not allow the full score. Please tag your problem set as "bmi713_pset5" on git.

If you have any questions, please post on the piazza site. This problem set was prepared by Tiziana Sanavia and Giorgio Melloni, so they will be most prepared to answer questions.

1. Multiple Regression models (30 points)

1.1 In this question we consider the meaning of the p-values of a linear regression. For this, we use the data set `punting`. A description of this data set can be found at http://www.statsci.org/data/general/punting.html. The data can be imported directly by:

```
read.table("http://www.statsci.org/data/general/punting.txt", header = TRUE)
```

- Perform a multiple regression of `Distance` on `R_Strength` and `L_Strength`. Compare the p-values for the coefficients and the p-value for the global F-test. (5 points)
- Perform now a simple linear regression of `Distance` on `R_Strength` and one of `Distance` on `L_Strength`. What are the p-values of the coefficients? (5 points)
- Compare the p-values for the coefficients from the multiple regression model and those obtained from the simple linear regression. Are they significant in both cases? If not, explain why this happens. (5 points)

1.2 The dataset `whiteside` is a collection of temperature and gas consumption of a UK house in the 60's. Temperatures and gas consumptions are recorded for two seasons, before and after a cavity-wall insulation was installed:

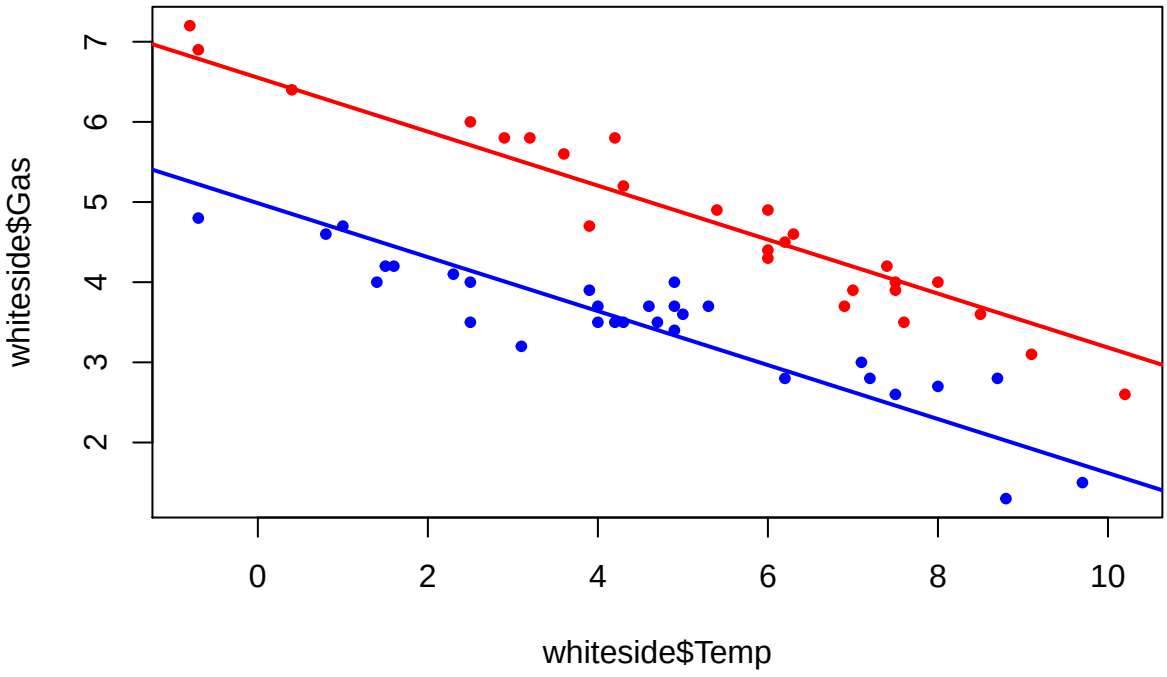
- `Insul` is a factor (Before and After insulation)
- `Temp` is a set of temperature in Celsius degrees
- `Gas` is gas consumption in cubic feet

The data set is available in the R package `MASS`:

```
library(MASS)
help("whiteside")
head(whiteside)
```

```
##    Insul Temp Gas
## 1 Before -0.8 7.2
## 2 Before -0.7 6.9
## 3 Before  0.4 6.4
## 4 Before  2.5 6.0
## 5 Before  2.9 5.8
## 6 Before  3.2 5.8
```

- Create a linear regression model considering the gas consumption depending on both insulation and temperature. Can these two variables explain the gas consumption? Interpret the coefficients obtained and their significance (one is a dummy variable the other is continuous!). Comment the results. (5 points)
- Display a scatter plot of temperature and gas with dots colored by insulation values. Add 2 regression lines for before and after insulation installation. (5 points)



- Create an interaction model to check whether there is a combined effect of insulation and temperature and comment the results obtained. (5 points)

2. Survival Analysis (10 points)

In the effort to determine whether two drugs used in treatment of thyroid disorders differed in terms of increasing the risk of cancer, researchers at Cambridge performed a study in which rats were randomly assigned to receive one of the drugs. The rats were then exposed to a known carcinogen, and the time until each rat died of cancer was recorded. The first few outcomes for one of the groups of rats is given below. There were 21 rats in this group.

Day	Numb_Deaths	Live_at_that_day
142	1	
156	1	
163	1	
198	1	
204	0	
205	1	
...	...	

- Complete the third column of the table, reporting the number $n(t)$ of rats alive at the beginning of the day. At the start of day 205, how many rats were at risk for dying of cancer? (3 points)

- What was the observed probability of dying from cancer at day 205? (3 points)
- Estimate the observed probability of surviving until day 160. (4 points)

3. Kaplan-Meier Survival Curves and the Log-Rank Test (20 points)

The data set "vets.csv" considers survival times in days for 137 patients from the Veterans Administration Lung Cancer Trial cited by Kalbfleisch and Prentice in their text (The Statistical Analysis of Survival Time Data, John Wiley, pp. 223-224, 1980):

- Column 1 = `treatment` (1 = standard, 2 = test)
- Column 2 = `cell type 1` (1 = large, 0 = other)
- Column 3 = `cell type 2` (1 = adeno, 0 = other)
- Column 4 = `cell type 3` (1 = small, 0 = other)
- Column 5 = `cell type 4` (1 = squamous, 0 = other)
- Column 6 = `survival time` (days)
- Column 7 = `performance status` (0 = worst, . . . , 100 = best)
- Column 8 = `disease duration` (months)
- Column 9 = `age` (years)
- Column 10 = `prior therapy` (0 = none, 10 = some)
- Column 11 = `status` (0 = censored, 1 = died)

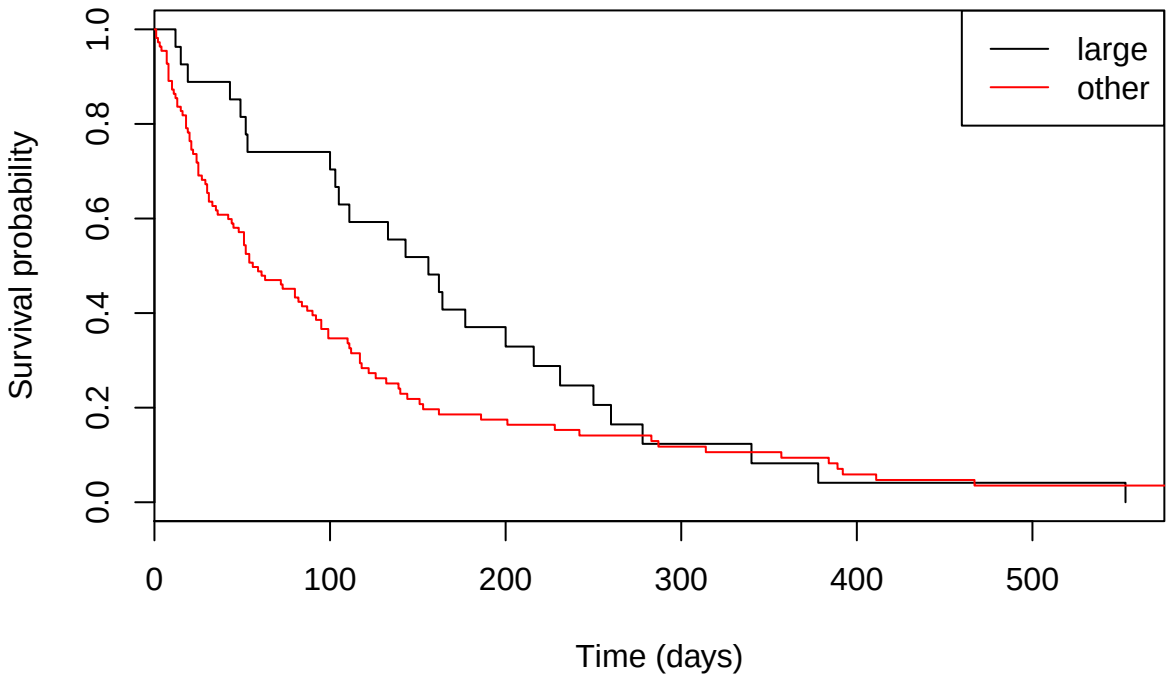
The data set can be dowloaded at the link

https://www.dropbox.com/sh/rlj714terrzakc7/AACELk752w4GPAsJXYi8Vbb5a?dl=0

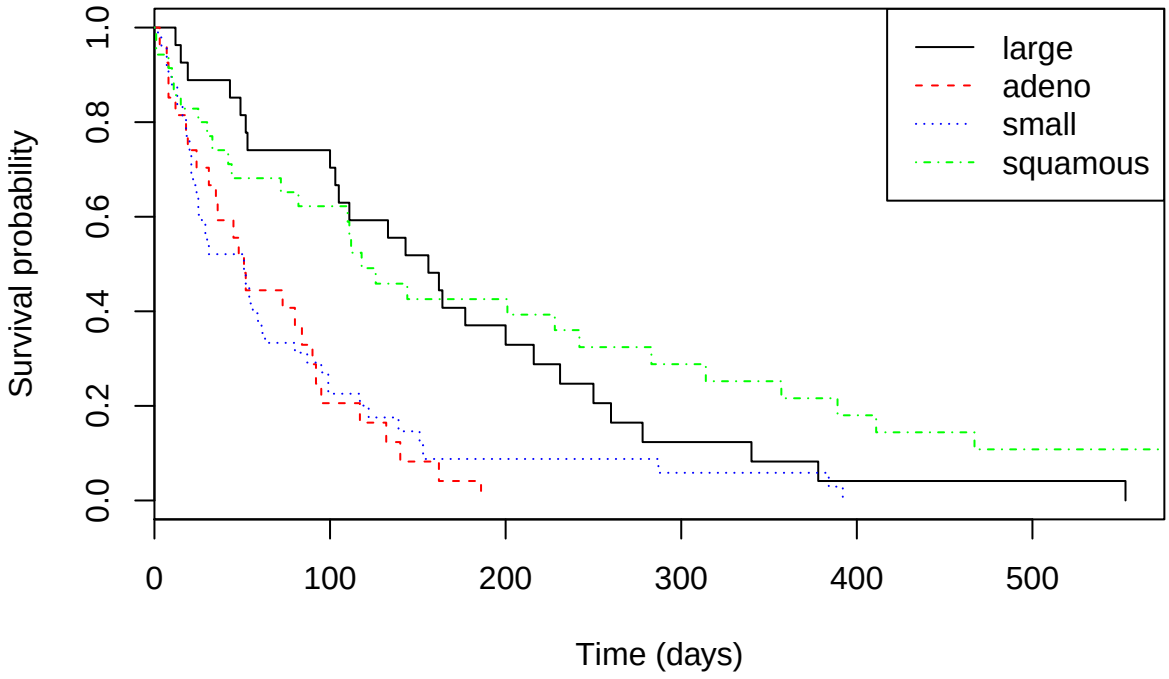
and then loaded in R using the command:

```
data <- read.csv("vets.csv")
```

- Obtain Kaplan-Meier plots for the two categories of the variable `cell type 1` (1 = large, 0 = other). Comment on how the two curves compare with each other. Moreover, carry out the log-rank test and draw conclusions from the test. NOTE: use function `survdiff` to perform the log-rank test. (10 points)



- Obtain Kaplan-Meier plots for the four categories of cell type-large, adeno, small, and squamous. Note that you will need to recode the data to define a single variable which numerically distinguishes the four categories (e.g., 1 = large, 2 = adeno, etc.). As in the previous part, compare the four Kaplan-Meier curves. Also, carry out the log-rank for the equality of the four curves and draw conclusions. NOTE: use function `survdiff` to perform the log-rank test. (10 points)



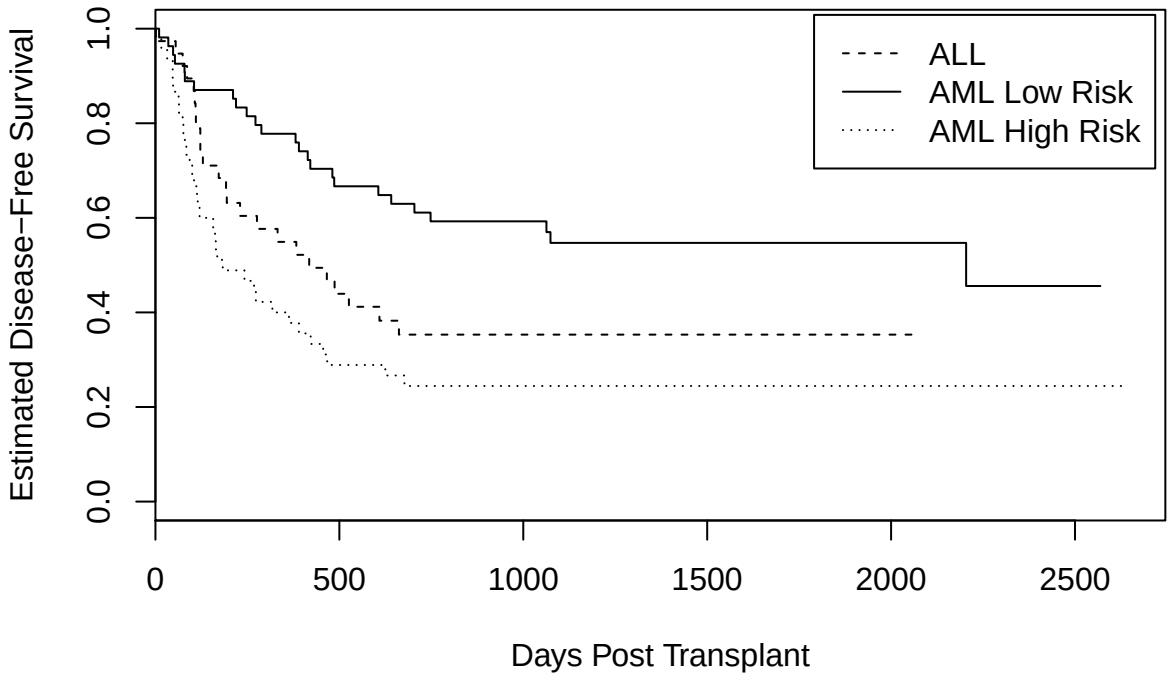
4. Cox models (40 points)

For the whole exercise we are investigating a dataset from a study by Copelan et al. (1991) of allogeneic marrow transplants for patients with acute myelocytic leukemia (AML) and acute lymphoblastic leukemia (ALL). A total of 137 patients (99 AML, 38 ALL) were treated at four different hospitals: 76 at the Ohio State University Hospital, 21 at Hahnemann University, 23 at St. Vincent's Hospital and 17 at Alfred Hospital. The data set is available in the R package `KMsurv`:

```
library(KMsurv)
data(bmt)
help(bmt)
bmt$group <- as.factor(bmt$group)
```

The last line is necessary for R since `group` has three levels. NOTE: Since `group` has three levels, the first level is treated as the base level and all remaining levels are compared with the base level. Therefore when it is tested in a model, there are two p-values provided for `group`. If one level of a categorical variable is significant then the full variable is deemed significant.

- Investigate the disease-free survival time using the Kaplan-Meier estimate for the survival curves. Plot the curves and comment briefly on what you see. What is the three year disease-free survival of the three patient types? HINT: The variables for the `Surv` function are $t2$ and $d3$ since we are interested in the disease free survival. (5 points)



- Check whether these three curves are really different using the log-rank test. Comment the results. NOTE: use function `survdiff` to perform the log-rank test. (5 points)
- We want to extend the basic comparisons of the three patient groups using a Cox PH model for time-independent covariates. Generate the Cox model using `group` as independent variable. Does the group 'AML Low Risk' (i.e. `group` = 2) reduce the hazard? If yes, by how much (report as a percentage)? (6 points)
- Test the model by adding confounding factors, using the following approach:

```
mod_var1 <- coxph(Surv(t2, d3)~group+var1, data=bmt)
summary(mod_var1)
```

Select the covariates according to their corresponding Wald test, testing each of the following covariates individually:

- var1 = z7 (waiting time to transplant)
- var1 = z8 (FAB class)
- var1 = z10 (MTX)
- var1 = z1 and z2 (patient and donor age)
- var1 = z3 and z4 (patient and donor gender)
- var1 = z5 and z6 (patient and donor CMV status)

For the variables age (z1, z2), gender (z3, z4) and CMV status (z5, z6) we want to use also the interaction term between donor and patient (HINT: use the $*$ symbol in the model, e.g. $group + z1 * z2$ is equivalent to $group + z1 + z2 + z1 : z2$).(5 points)

- Look at the results and then consider the model with the linear combination of `group` and all the resulting significant covariates. Considering this latter model then try to:

(a) replace one of the significant covariate with each of the not-significant covariate at a time and look at the performance of the model (Wald test). Example: if z7 and $z3 * z4$ are the only significant covariates, try to replace $z3 * z4$ first with z8 and check the results of the model, then with $z1 * z2 \ldots$. (5 points)

(b) keep all the significant covariates and add each of the not-significant covariate at a time and look at the performance of the model (Wald test). Example: if z7 and $z3 * z4$ are the only significant covariates, try to use z7, z8 and $z3 * z4$ and check the results of the model, then try z7, $z3 * z4$ and $z1 * z2 \ldots$ (5 points)

Are there combinations tested in (a) and (b) better than the model that combines only the covariates which resulted significant when they were individually used in the original model with `group`? (3 points)

- Use stepwise selection (in both directions, i.e. backward and forward) using the function `step` which uses the AIC (Akaike information criterion) and compare the result with the best model previously obtained. Look at the final model obtained from the stepwise selection: which covariate shows the strongest association with poor survival? (Provide a proper explanation to your choice) (6 points)

5. Extra: Multiple testing (5 points)

Create an R function which receives as input the p-values and provides as output the highest p-value with a corresponding Benjamini-Hochberg FDR < 0.05. Do not use or copy the implementation provided by the `p.adjust` function. HINT: the usage of `while` can be useful.