Lecture 2: October 31, 2017

Lecturer: Prof. Ron Shamir *Scribe: Kathy Razmadze*

2.1 Multiple testing and FDR

2.1.1 Hypothesis Testing

In this part we give a brief primer on statistical and analytical tools that will be needed later in the course. For sources e.g. [4] [6]
Suppose there are two opinions for the distribution from which a sample was taken:
- H_0 - the *null hypothesis*. This is our default assumption.
- H_1 - the *alternative hypothesis*
The assumptions are not symmetric.The null hypothesis is assumed unless we have strong evidance to reject it. We will need a test that has a range C of values of the sample outcome for which the null assumption is rejected. C is called the *rejection region*. We can summarize all the options by Table 2.1.

We will be more concerned to have a mistake of Type 1, because in that case we will reject

Table 2.1: Assumption options

		decision	
		Reject H_0	Accept H_0
Truth	H_0 True	Type 1 error	correct
Truth	H_1 True	correct	Type 2 error

assumption H_0 which is the preferred assumption. Assume, for example, that there is a test that will tell whether a patient has a certain life risking disease (H_0). If we will conclude that the patient has a disease, an aggressive treatment will be given to the patient. If not (H_1), the patient will be diagnosed with no disease and be discharged. If this diagnosis was wrong - he might die. Hence a type 1 error has more severe consequences than type 2 errors. Formally define:

$\alpha = Pr(Type\ 1\ error) = Pr_{H_0}(C) = Pr(C|H_0\ true)$
$\beta = Pr(Type\ 2\ error) = Pr_{H_1}(\neg C)$
$\pi = Pr(Reject\ H_0\ when\ H_1\ is\ true) = Pr(C|H_1\ true) = 1 - Pr_{H1}(\neg C) = 1 - \beta$

The power of test should be maximized to get the best results, and it is called the *power* of the test, and α is called the statistical significance. The test subject to an upper bound on α. We would like a test that keeps type 1 error smaller than α and minimizes β. In other words, we want a test with maximum power as long as the probability for type 1 error smaller then α.

Assume there is a sample X = $\{X_1, ..., X_n\}$ from distribution p, denote P_0(x) as the probability of the sample under H_0. In the same way, denote P_1(x) as the probability of x under H_1.
Define *likelihood ratio* as: λ (X) = P_1(X) / P_0 (X)

Theorem 2.1 *Neyman-Pearson Lemma : For single hypothesis, a test with maximum power has the form C =* $\{\ \lambda(X) > K\ \}$ *where K is set to get significance* α.

(A hypotheseis is *simple* if it uniquely determinea the distribution.) By the lemma we can sort the possible outcomes in decreasing order of λ and put in C all outcomes as long as α is not exceeded.
Assume a sample test result y. Denote p = Prob (getting result y or more extreme | H_0). We will reject H_0 if p$\leq \alpha$. It is common to use $\alpha = 0.05$.

2.1.2 The Permutation test

Suppose there are two samples denoted X = $\{X_1, ..., X_m\}$ and Y= $\{Y_1, .., Y_n\}$ samples from distributions F_1 and F_2 respectively. There are two alternatives:

$$H_0 = F_2 = F_1$$
$$H_1 = F_2 \neq F_1$$

(For example, assume that blood pressure masurements were taken from the members of two groups as for treatment. One group was given a new medical treatment and the second got a placebo. The goal of the test is to check whether their distributions are equal or not, i.e., whether the medication is beneficial)

A possible statistic T is: T($X_1, ..., X_m$,$Y_1, ..., Y_n$)=| Ave(X)-Ave(Y) |. Assume rejection for high T (large difference). Suppose T=t_{obs} on the sample, Is this value significant?
To check that, there are 2 options:
Consider all the permutations on the N=n+m elements in the sample, and for each one of them compute T_i= | Ave(X)-Ave(Y) |. We assume that all the permutations have equal probability mass $1/N!$ on each T_i. This assumption is called the *permutation distribution.* Now count the number K of permutations that get a value bigger than t_{obs}. Then p = $\frac{K}{N!}$ (p-val = fraction of i for which $T_i > t_{obs}$). Ee will compare p with α and then decide whether to reject H_0.
The problem with this option is that N is usually too large. Therefore, the second option is to generate a reasonable number L of random permutations. In this case an approximate/empirical p-val is received, but it is limited to values above 1/L is possible between values $\leq$ 1/L.

2.1.3 Multiple testing problem

Suppose we preform a test on multiple genes, having a seperate hypothesis for each one. For example, we measure expression levels of 10K genes in 10 individuals, 5 of them are sick in a disease (they are referred to as *cases*) and 5 healthy (*controls*). To check whether the level of gene number 1 is different between the groups we will use the T statistic defined above. Because there are 10K such tests, the chance of at least one false rejection is much higher than in just one test. More generally, suppose we have m hypotheses tested: H_{0i} compared to H_{1i} for $i = 1, .., m$. The p-values obtained will be denoted as p_1,..,p_m. [6]

2.1.4 Bonferroni Method

Bonfferoni's method rejects null hypothesis H_{0i} if $p_i < \alpha$/m, for $i = 1, .., m$.

Theorem 2.2 *The probabilty of falsely rejecting even one null hypothesis is smaller than* α.

Proof:

Denote by R_i the event that the i-th hypothesis is falsely rejected. We now compute the probability that we will have at least one false rejection:
P ($\geq$1 false rejection) = P ($\cup_{1..m} R_i$) $\leq \Sigma_i$ P (R_i) =Σ_i (α/m) = α ■

Note that the method makes no assumption on the distribution. This method is highly conservative, as it aims to avoid even one false positive error. As a result we lose power.

2.1.5 Benjamini-Hochberg Method

An alternative approach to multiple hypothesis is using false discovery rate. suppose our rejection rule gives the outcomes as shown in table 2.2.

Table 2.2: Number of outcomes

		Decision		
		*Reject*H0	Accept H_0	Total
Truth	H_0 True	V	U	m_0
Truth	H_1 True	S	T	m_1
	Total	R	m-R	m

Denote FD proportion (FDP) as the fraction of rejections that are incorrect, i.e.FDP = V/R (0 if R=0). Define the *False disocery rate* as FDR = E(FDP)
The method of Benjamini and Hochberg works as followes:
Reorder the hypotheses $(H_1, ..H_m)$ such that $p_1 < p_2 < .... < p_m$. Find a maximum value k such that $p_k < k\alpha/m$. Reject the null hypotheses $(H_1, ..H_k)$

Theorem 2.3 *[Benjamini and Hochberg 1995 [1]]If the p-values are independent, the procedure guarantees that FDR* $< \alpha$

See Figure 2.1 for an example.Example define α=0.05, and assume there are 1000 hypotheses. Bonferroni threshold: 0.00005. when using FDR, a hypothesis 20, 20 x 0.05/1000= 0.001. Hence if k=2 out of the first 20 hypotheses, we expect one false rejection.

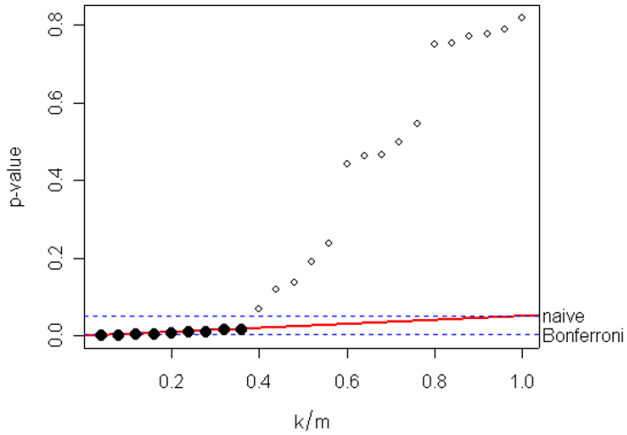


Figure 2.1: BH example, the hypothesis in black are rejected

We can look a bit differently on the corrections. Instead of adjusting the threshold, the p-values can be adjusted to check if they are below the fixed threshold. Denote the adjusted p-values by $p*_i$. Then the adujsted p-values are:
Bonferroni: $p*_i = p_i m$
FDR: (after sorting the $p_i - values$)$p*_i = p_i m/i$
In both cases we reject if $p*_i < \alpha$

2.2 Survival Analysis

2.2.1 Introduction and Terminology

Figure 2.2, is one of the first survival curves that was evercreated. The curve shows how many people survived till the age of X (for example, 16 people survived till the age of 36)[5]

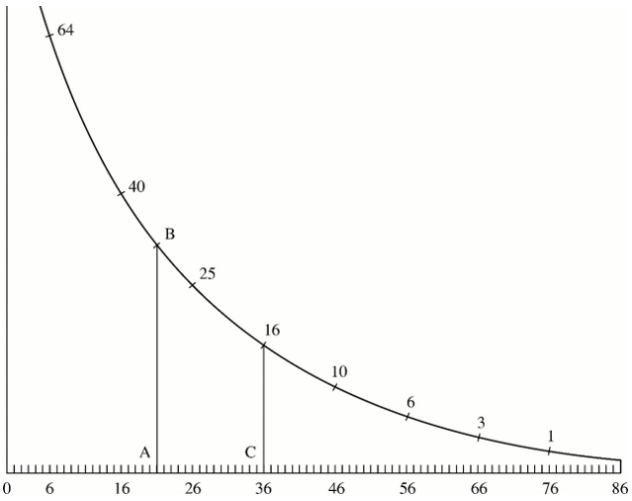
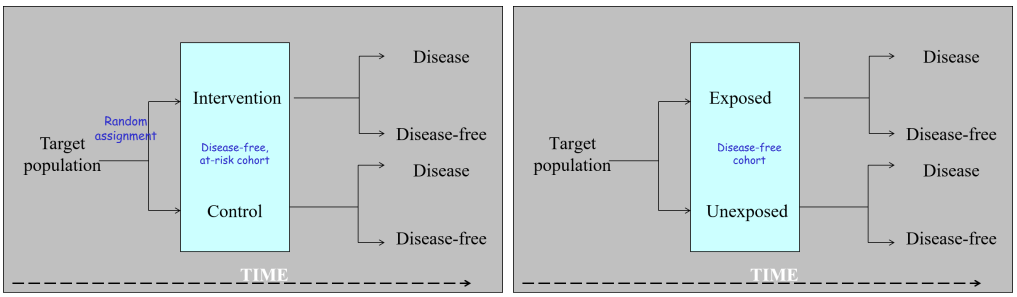


Figure 2.2: Christiaan Huygens' 1669 curve showing how many out of 100 people survive until 86 years

Survival analysis is a branch of statistics that analyzes longitudinal data on the occurrence of event. The event is predefined and is followed up and measured. Such analysis can be made on several possible events, like death, injury, onset of disease, recovery from illness, and recurrence-free survival for 5 years, (binary variables). The event could also be a transition above or below the clinical threshold of a continuous variable (e.g. blood glucose level). Rejecting event is binary in all cases. The data for the analysis can be obtained from a randomized clinical trial or a cohort study design.[3]

In a *Randomized Clinical Trial* (RCT) a group of people is randomly divided into two subgroups. One group is treated (e.g given a medication), while the other group is a control group (e.g. given a placebo) (figure 2.3):



(a) Check if the intervention cases the disease

(b) cohort study - check if the exposure cause a disease

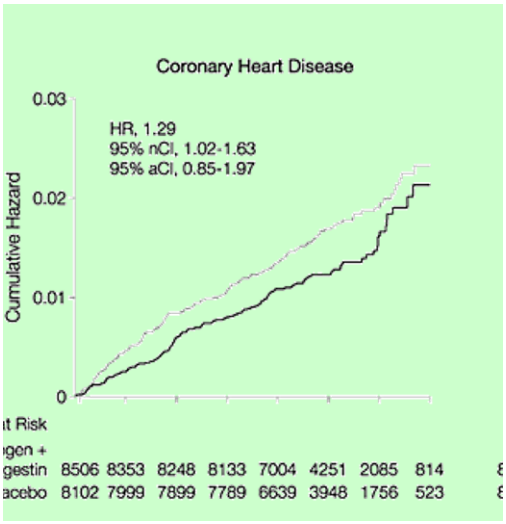
Figure 2.3: An RCT (left) and a cohort study (right).

A *cohort study* takes a group of people, partitions them into two groups based on having a disease, and checks the survival of each group. It could be *prospective*, watching the patients from the moment we have started our experiment. It could be *retrospective*, where we have already gathered the data on a group of patients and we know if they have been exposed or not. The advantage in the prospective way is that we can define all parameters before the experiment has started. On the other hand, in retrospective analysis the data could be biased. There are many examples of survival analysis in medicine, one of them is the chance that women that received hormones will have a coronary heart disease. Another one is the connection between low-fat diet and breast cancer.(Figure 2.4) [2]

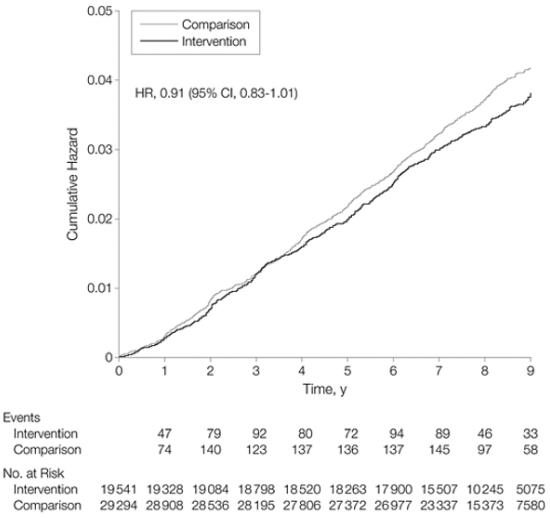
Survival analysis has several advantages over other methods, for example comparing mean time-to-event between groups using a t-test or linear regression: For some patients we may not know if and when an event occurred, because the study terminated, or we lost touch with them. It is also problematic to compare proportion of events in each group using risk, odds ratios or logistic regression, because we ignore the time dimension.

We introduce several definitions of the terminology we shall use. The *event* of interest is the outcome sought. The *Time-to-event* is the time from entry into a study until a subject had the outcome. Subjects are said to be *censored* if they are lost to follow up or drop out of the study, or if the study ends before they have the outcome. They are counted as alive / disease-free for the time they were enrolled in the study. Censoring must be independent of the outcome, otherwise censoring will create bias.

Two-variable outcome :
t_i = time at last disease-free observation or time at event



(a) The connection between hormones and coronary heart disease



(b) The connection between breast cancer and low-fat diet

Figure 2.4: Different types of examples to RCT

$c_i = 1$ if had the event; $c_i = 0$ if no event happened by time t_i

2.2.2 Survival/hazard functions

Define S(t) to be the probability of an individual surviving at least until time t. S(t) is usually unknown, and can be evaluated based on a sample. The term *survival experience* is sometime used for the empirical function.

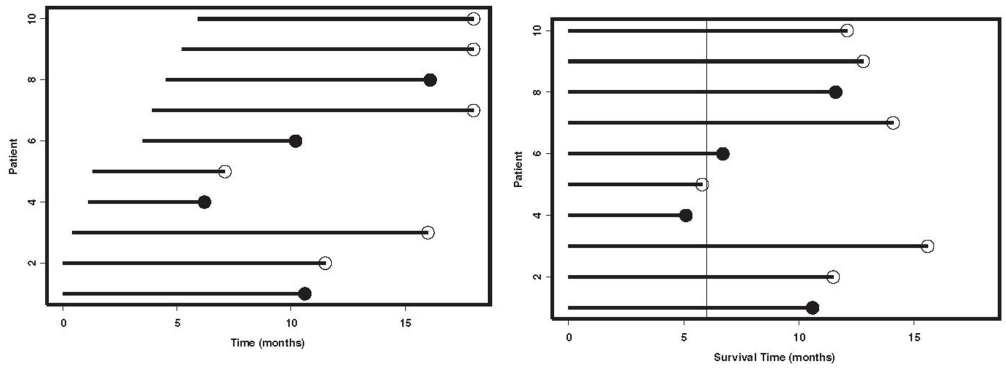
Let T be random variable indicativing the event time for an individual. The probability of the event time occurring at exactly time t is

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}$$

We denote f by F(t) the CDF of f(t) (the probability of dying by time t)
Then, S(t) = 1-F(t)
The *hazard* function f, is the probability that if you survive to t, you will succumb to the event in the next instant. Formally:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t / T \geq t)}{\Delta t}$$

See Figure 2.7. Using Bayer's rule we get:



Patient No.	Time at Entry (Months)	Time at Death or Censoring (Months)	Dead (D) or Censored (C)	Survival Time (Months)
1	0.0	10.6	D	10.6
2	0.0	11.5	C	11.5
3	0.4	16.0	C	15.6
4	1.1	6.2	D	5.1
5	1.3	7.1	C	5.8
6	3.5	10.2	D	6.7
7	3.9	18.0	C	14.1
8	4.5	16.1	D	11.6
9	5.2	18.0	C	12.8
10	5.9	18.0	C	12.1

Figure 2.5: Survival data and censoring. left: time from entry to death (solid circles) or censoring (open) circles. Middle: The same data after moving all start times to 0. Right: The same data in tablular form

$$h(t)dt = P(t \leq T < t+dt/T \geq t) = \frac{P(t \leq T < t+dt \& T \geq t)}{P(T \geq t)} = \frac{P(t \leq T < t+dt)}{P(T \geq t)} = \frac{f(t)dt}{S(t)}$$

Hence we obtained the following relation between hazard, density and survival

$$h(t) = \frac{f(t)}{S(t)}$$

2.2.3 Kaplan-Meier curves

When the events are distinct and sorted $t_1 < t_2 < ... < t_n$ and there is no censoring, then

$$Pr(surviving\ to\ t_i) = \frac{n-i+1}{n}$$

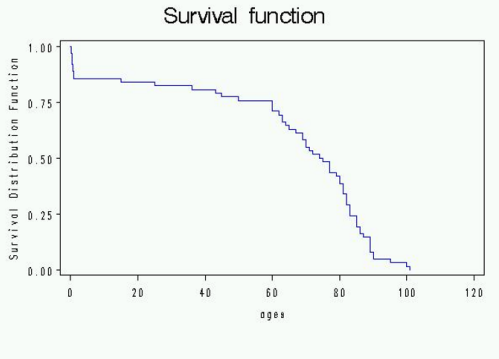


Figure 2.6: Comulative survival

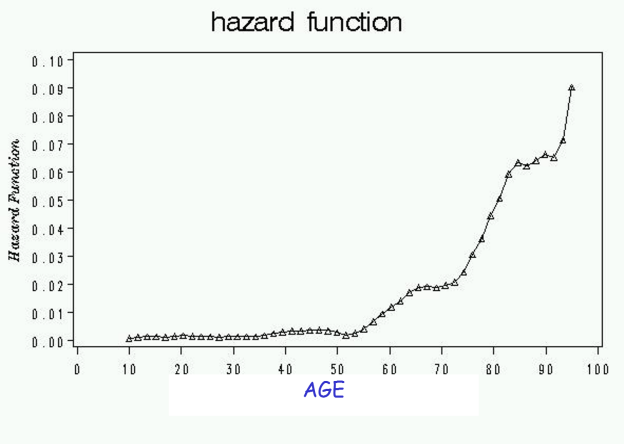


Figure 2.7: Hazard function

Suppose some subjects are censored, and the events are sorted $t_1 < t_2 < ... < t_n$ and there are d_i be the events in the interval $(t_{i-1}, t_i]$. Let n_i be the number of individuals *at risk* (the people that remain in the study) in the time slot $(t_{i-1}, t_i]$
And then we get:
Pr(survival to t_i)= P(surviving to t_{i-1}) x P(surviving interval $(t_{i-1}, t_i]$) = $P(survival\ to\ t_{i-1})\frac{n_i - d_i}{n_i}$
Which results gives:

$$\tilde{S}(t) = \Pi_{i|t_i \leq t}(1 - \frac{d_i}{n_i})$$

Figure 2.8 and 2.9 present an example of censored cumulative survival function. The plot $\tilde{S}$ is called *Kaplan-Meier plot*. Bars indicate times of censoring. When a person is censored no event is counted but the number of the persons at risk decreases. The censored persons do not affect the shape of the Kaplan-Meir plot. For example, at time 6 there is one censoring

(marked by circle) and no event, that is why the proportion of survival stays the same.
The Kaplan Meier estimate and curve is a non-parametric estimate of the survival function.

Time, t_i	No. at Risk, r_i	No. of Events, f_i	Product-Limit Estimator	Cumulative No. Censored, $m_c(t_i)$	Cumulative No. of Events, $m_e(t_i)$
0	20	0	1.00	0	0
5	20	2	$1 - 2/20 = 0.90$	0	2
6	18	0	$(1 - 0/18) \times 0.90 = 0.90$	0	2
10	15	1	$(1 - 1/15) \times 0.90 = 0.84$	3	3
13	14	2	$(1 - 2/14) \times 0.84 = 0.72$	3	5

Figure 2.8

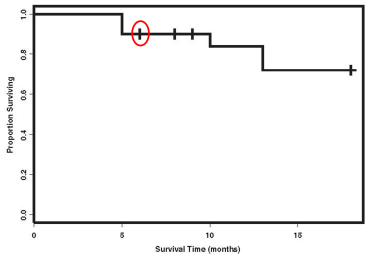


Figure 2.9: Cumulative survival function: Kaplan - Meier Curve

It is the empirical probability of surviving up to end time point in the sample, taking into account censoring. It describes survivorship of study population/s, and commonly used to compare two study populations. As you may see in the example, it has a intuitive graphical presentation.

2.2.4 The LogRank test

In order to compare two survival curves, one can compare the curves at a pre-specified time point t. The result depends on t, and determining the right t a-priori is difficult.
Alternativly, we can compare the overall plots over the entire time range. This is what we shall discuss here.

The *Log rank test* evaluate the hypothesis $H_0 : S_1(t) = S_2(t)$ for all t. if uses the ranks of events instead of the times.Sort the events $t_1 < t_2 < ... < t_K$. For time t_j (interval $(t_{j-1}, t_j]$) the number of events of each group is presented in Table 2.3.

Under H_0, $E(a_j)$= (total number of events) x (number of people at risk group 1)/ (number of people at risk) = $\frac{(a_j+c_j)(a_j+b_j)}{n_j}$

Table 2.3: Number of persons at risk and events of t_j

	Events	**Surviving**	**Total**
Group1	a_j	b_j	$a_j + b_j$
Group2	c_j	d_j	$c_j + d_j$
Total	$a_j + c_j$	$b_j + d_j$	n_j

Since Z is approximately standard normal we can evaluate it's p-values.

$$Var(a_j) = \frac{(a_j + b_j)(a_j + c_j)(b_j + d_j)(c_j + d_j)}{(n_j - 1)(n_j)^2}$$

$$Z = \frac{\sum_{j=1}^{k} a_j - E(a_j)}{\sqrt{\sum_{j=1}^{k} Var(a_j)}}$$

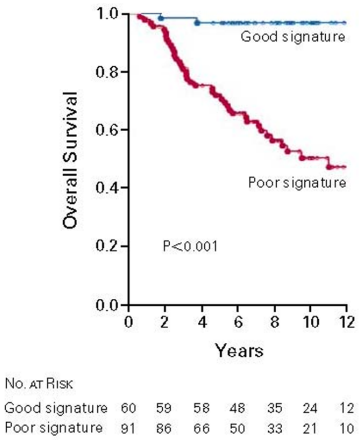
Several comments must be made about KM curve and logRank. First, as there is censoring it makes no sense to talk about mean survival. Second, visual inspection can be misleading. For example, at the right side of Figure 2.17 you can see that the number of persons at the experiments is much smaller). One must also predefine the groups in advance (and not after we have seen the results).

In addition, the parameters are binary. Certain characteristics such as age, sex, etc. are not taken in account, while they can be related to survival. Such *confounding / prognostic factors* can change the relation of treatment to outcome. To account for confounding we need to stratify the test and compare survival differences within each level of these factors However, the resulting groups may be small and we lose power.

2.2.5 Cox Proportional Hazard

Now, we will describe the Cox Proportional Hazard model. Kaplan-Meir curves and Log Rank are univariate analysis, as they describe survival using one categorical factor. Cox PH allows many prognostic factors, categorical or real-valued. Cox PH is semi-parametric, and can model the effect of predictors and covariates on the hazard rate but leaves the baseline hazard rate unspecified. It estimates relative rather than absolute hazard.
The model assuming a *baseline hazard function* $\lambda_0(t)$ that is left unspecified but must be positive. It is the hazard when all covariates are 0. In addition, a linear function of a set of k fixed covariates that is exponential.

$$h_i(t) = \lambda_0(t) e^{\beta_1 x_{i1} + \ldots + \beta_k x_{ik}}$$



sig.png

Figure 2.10: Example: breast cancer survival signature

Or equivalently

$$log h_i(t) = log\lambda_0(t) + \beta_1 x_{i1} + ... + \beta_k x_{ik}$$

We will define Hazard ratio as follows:

$$HR_{i,j} = \frac{h_i(t)}{h_j(t)} = \frac{\lambda_0(t)e^{\beta_1 x_{i1}+...+\beta_k x_{ik}}}{\lambda_0(t)e^{\beta_1 x_{j1}+...+\beta_k x_{jk}}} = e^{\beta_1(x_{i1}-x_{j1})+...+\beta_1(x_{ik}-x_{jk})}$$

By assumption, hazard functions should be strictly parallel and produce covariate-adjusted hazard ratios.
The point is to compare the hazard rates of individuals who have different covariates, hence, called Proportional hazards. Note that $\lambda_0(t)$ cancelled out. For example, $h_i(t)$ could be the hazard for person i, that is a smoker, and $h_j(t)$ could be hazard for person j, that is a non-smoker. Then $HR_{i,j}$ gives the hazard rate for smoking.

$$HR = \frac{h_1(t)}{h_2(t)} = \frac{h_0(t)e^(\beta x_1)}{h_0(t)e^(\beta x_2)} = e^(\beta(x_1 - x_2))$$

The coefficients $\beta_1, , \beta_K$ can be estimated using numerical optimization (details not shown). For large enough sample, the estimate of each β_i has a normal distribution and its p-val and confidence intervals can be computed.

Example: the Farmingham heart study, is a cohort of 5,180 persons aged 45-82 who were followed until time of death or up to 10 years. 46 percent are males, and there were 402 deaths. Table 2.4 summarize the results.

Cox PH model for age and sex as factors is presented in Table 2.5.

The analysis shows that both factors increase risk. For age, exp(0.11149) = 1.118 that is 11.8 percent higher risk per year. For gender, Male, exp(0.67958) = 1.973, so there is 97.3

Table 2.4: The Farmingham results

	Die (n=402)	**Not Die (n=4778)**
Mean (SD) Age, Years	65.6(8.7)	56.1(7.5)
N (percent) Male	221 (55 percent)	2145 (45 percent)

Table 2.5: Cox PH analysis for age and sex

Risk factor	**Parameter Estimate**	**P-Value**
Age, Years	0.11149	0.0001
Male Sex	0.67958	0.0001

precent higher risk per males, holding age constant

Figure 2.11 shows the cox PH results for a model with more covariates:
The four first factors are significant and have CIs that do not include 1 (the null).

Risk Factor	Parameter Estimate	P-Value	Hazard Ratio (HR) (95% CI for HR)
Age, years	0.11691	0.0001	1.124 (1.111-1.138)
Male Sex	0.40359	0.0002	1.497 (1.215-1.845)
Systolic Blood Pressure	0.01645	0.0001	1.017 (1.012-1.021)
Current Smoker	0.76798	0.0001	2.155 (1.758-2.643)
Total Serum Cholesterol	-0.00209	0.0963	0.998 (0.995-2.643)
Diabetes	-0.02366	0.1585	0.816 (0.615-1.083)

Figure 2.11: Cox PH results: Taking in account age, sex, blood pressure, smoking, cholesterol, diabetes

Bibliography

[1] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Royal Statistical Society*, 1995.

[2] Aruna D. Inflammatory biomarkers, hormone replacement therapy, and incident coronary heart disease. *JAMMA*, 288.

[3] Rich Jason T. A practical guide to understanding kaplan-meier curves. 2010.

[4] T. Livyatan and A.Raviv. Introduction to probability and statistic. 2005.

[5] Howard Wainer. Statistical graphics: Mapping the pathways of science. *Annual Review of Psychology*, 52.

[6] Larry A. Wasserman. All of statistics: A concise course in statistical inference. *Springer*, 2004.