Lecture 10:
Cox Model & Multiple Testing

BMI 713
November 21, 2017
Peter J Park

Cox PH Model

- We are often interested in the relationship between survival time and a continuous risk factor, or to evaluate the simultaneous effects of more than one risk factor
- Log-rank test is only for one dichotomous variable
- Multivariable analysis can be performed using the **Cox proportional hazards model**
- Multiple linear regression analysis cannot be used because survival time is rarely normally distributed, and because it cannot account for censored observations
- The Cox model is an example of a **semiparametric** model

Cox PH Model

- We need a new function called the **hazard function, *h(t)***
- This is the probability that you will die in the very instant after time *t*, given that you have survived until time *t*
- The proportional-hazards model assumes that the hazard rate for any individual can be modeled as a function of covariates X_1, …, X_k as follows:

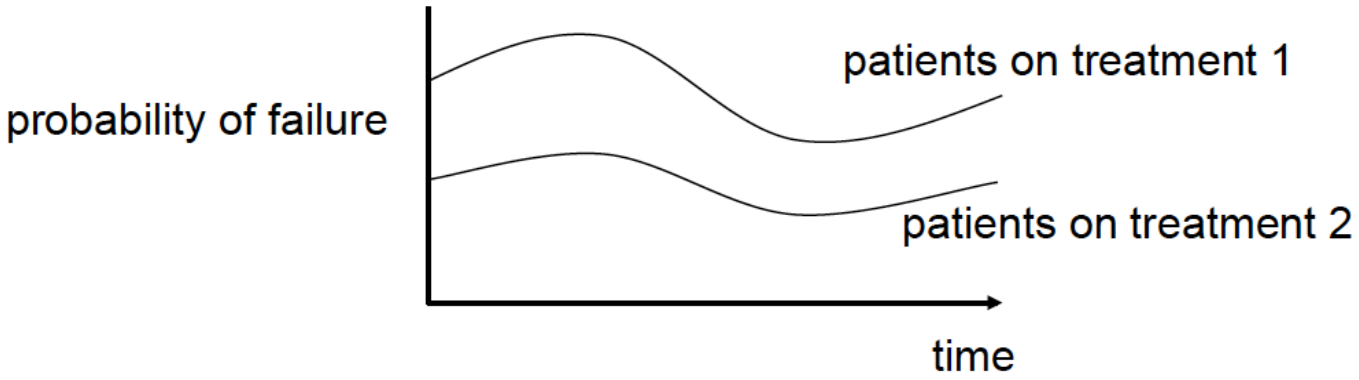
$$h(t) = h_0(t)e^{\beta_1 x_1 + \cdots + \beta_k x_k}$$

$$\ln\left(\frac{h(t)}{h_0(t)}\right) = \beta_1 x_1 + \cdots + \beta_k x_k$$

Cox PH Model

$$h(t) = h_0(t)e^{\beta_1 x_1 + \cdots + \beta_k x_k}$$

- h0(t) is called the “baseline hazard rate”
- We make no assumptions about its shape
- This is why the model is called semi parametric. We don’t completely specify the distribution of survival times; we only specify that changes in covariates will change the hazard rate proportionally to whatever it was.



Interpretation of the Coefficients

- Interpreting the parameters of the model is a bit difficult. The easiest case to understand is when a variable is dichotomous.
- Example: Suppose we are analyzing survival times using a Cox PH model with covariates X_1 = gender (1=F), X_2 = drug dosage. What is the ratio of hazards between a man and a woman on the same dose of the drug?

$$\frac{h_{woman}(t)}{h_{man}(t)} = \frac{h_0(t)e^{\beta_1(1)+\beta_2 x_2}}{h_0(t)e^{\beta_1(0)+\beta_2 x_2}} = e^{\beta_1}$$

- β_1 is the logarithm of the **"hazard ratio"**, which can be thought of as the instantaneous relative risk of death per unit time of a woman vs. of a man, given that both have survived until time t and with all other covariates held constant

https://www.r-bloggers.com/cox-proportional-hazards-model/

Cox PH example

```
install.packages(c("survival", "survminer"))
library("survival")
library("survminer")
data("lung")
head(lung)
```

	inst	time	status	age	sex	ph.ecog	ph.karno	pat.karno	meal.cal	wt.loss
1	3	306	2	74	1	1	90	100	1175	NA
2	3	455	2	68	1	0	90	90	1225	15
3	3	1010	1	56	1	0	90	90	NA	15
4	5	210	2	57	1	1	90	60	1150	11
5	1	883	2	60	1	0	100	90	NA	0
6	12	1022	1	74	1	1	50	80	513	0

- inst: Institution code
- time: Survival time in days
- status: censoring status 1=censored, 2=dead
- age: Age in years
- sex: Male=1 Female=2
- ph.ecog: ECOG performance score (0=good 5=dead)
- ph.karno: Karnofsky performance score (bad=0-good=100) rated by physician
- pat.karno: Karnofsky performance score as rated by patient
- meal.cal: Calories consumed at meals
- wt.loss: Weight loss in last six months

Loprinzi et al.
Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group.
Journal of Clinical Oncology.
12(3):601-7, 1994

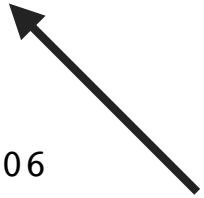
```
res.cox <- coxph(Surv(time, status) ~ sex, data = lung)

summary(res.cox)
Call:
coxph(formula = Surv(time, status) ~ sex, data = lung)

  n= 228, number of events= 165

       coef exp(coef) se(coef)      z Pr(>|z|)
sex -0.5310    0.5880   0.1672 -3.176  0.00149 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    exp(coef) exp(-coef) lower .95 upper .95
sex     0.588      1.701    0.4237     0.816

Concordance= 0.579  (se = 0.022 )
Rsquare= 0.046   (max possible= 0.999 )
Likelihood ratio test= 10.63  on 1 df,   p=0.001111
Wald test            = 10.09  on 1 df,   p=0.001491
Score (logrank) test = 10.33  on 1 df,   p=0.001312
```

- exp(coef) = exp(-0.53) = 0.59 is the hazard ratio (for the second group relative to the first). Being female (sex=2) reduces the hazard by a factor of 0.59, or 41%. Being female is associated with good prognostic.

Multiple Cox Regression

```
res.cox <- coxph(Surv(time, status) ~ age + sex + ph.ecog, data =  lung)
summary(res.cox)
Call:
coxph(formula = Surv(time, status) ~ age + sex + ph.ecog, data = lung)

  n= 227, number of events= 164
   (1 observation deleted due to missingness)

             coef exp(coef)  se(coef)      z Pr(>|z|)
age      0.011067  1.011128  0.009267  1.194 0.232416
sex     -0.552612  0.575445  0.167739 -3.294 0.000986 ***
ph.ecog  0.463728  1.589991  0.113577  4.083 4.45e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

        exp(coef) exp(-coef) lower .95 upper .95
age        1.0111     0.9890    0.9929    1.0297
sex        0.5754     1.7378    0.4142    0.7994
ph.ecog    1.5900     0.6289    1.2727    1.9864

Concordance= 0.637  (se = 0.026 )
Rsquare= 0.126   (max possible= 0.999 )
Likelihood ratio test= 30.5  on 3 df,   p=1.083e-06
Wald test            = 29.93  on 3 df,   p=1.428e-06
Score (logrank) test = 30.5  on 3 df,   p=1.083e-06
```

Being female (sex=2) reduces the hazard by a factor of 0.58, or 42%.



HR = 1.59, ph.ecog is associated with a poor survival.

Summary

- **Survival analysis** to handle survival data which usually have censored data points and are non-normally distributed
- **Kaplan-Meier estimator** for estimation & one-sample inference
- **Log-Rank test** for Two-sample comparisons
- **Cox Proportional Hazards model** for regression modeling

Regression Models - Summary

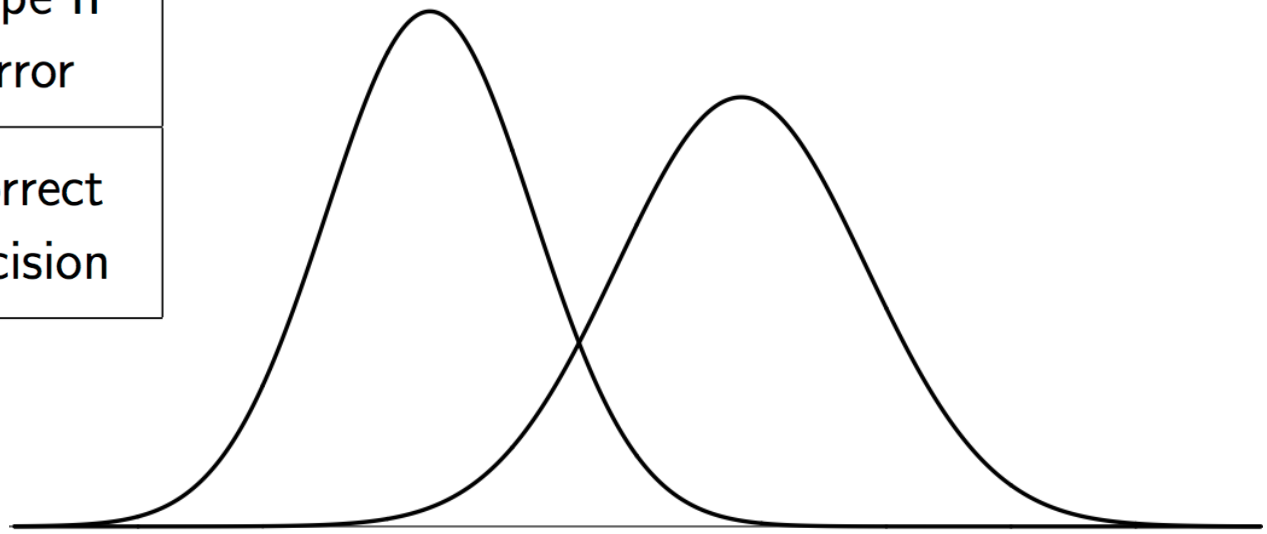
- Binary (disease vs. normal) ➔ Logistic regression (and many others!)
- Discrete
  - Non-ordered (multiple subclasses) ➔ Polytomous regression
  - Ordered (number of recurrences) ➔ Poisson regression

- Continuous (gene expression) ➔ Linear regression
- Censored (patient survival time) ➔ Cox model

Multiple Hypothesis Testing

- In many situations, there are many null hypotheses to test.
- You are bound to get false positives if you do not account for the fact that there are multiple hypotheses.

	H_0 is true	H_1 is true
Do not reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision



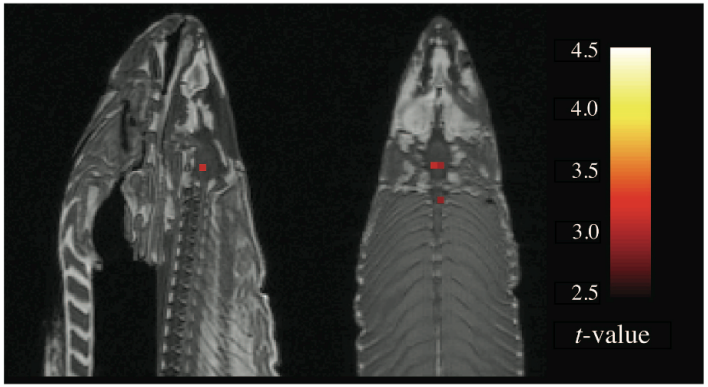
Multiple comparison correction

Winner of the 2012 IgNobel Prize in neuroscience

Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction

Craig M. Bennett [1]*, Abigail A. Baird [2], Michael B. Miller [1] and George L. Wolford [3]

[1]Department of Psychology, University of California at Santa Barbara, Santa Barbara, CA 93106
[2]Department of Psychology, Blodgett Hall, Vassar College, Poughkeepsie, NY 12604
[3]Department of Psychological and Brain Sciences, Moore Hall, Dartmouth College, Hanover, NH 03755



- The dead salmon was shown a series of photos
- fMRI images were taken before and after
- Three out of 130,000 voxels were significant ($p<0.001$) [when multiple testing is not used]

Where was the paper published?

What happened when you submitted the dead-salmon paper?

We tried to get it published in two major neuroimaging journals. One rejected it and the other sent it out for review. One reviewer said it was fantastic; the other gave us a hateful, livid review that sunk it. But less-mainstream journals were clamouring for the paper. We went with the *Journal of Serendipitous and Unexpected Results*, which led to other publications and fostered a debate on statistical errors.

Has the field changed?

In the salmon paper, we did a meta-analysis of major journal articles and found that 25–40% of neuroimaging papers that we studied were not properly correcting for threshold values. We surveyed a couple of journals last year as a follow-up, and found that fewer than 10% of people are now using incorrect statistics. The decline is not all attributable to the salmon paper, but it is all progress. We gave the field a kick in the pants — and I've heard that a lot of groups reviewed the paper in lab meetings.

http://www.nature.com/naturejobs/science/articles/10.1038/nj7420-437a

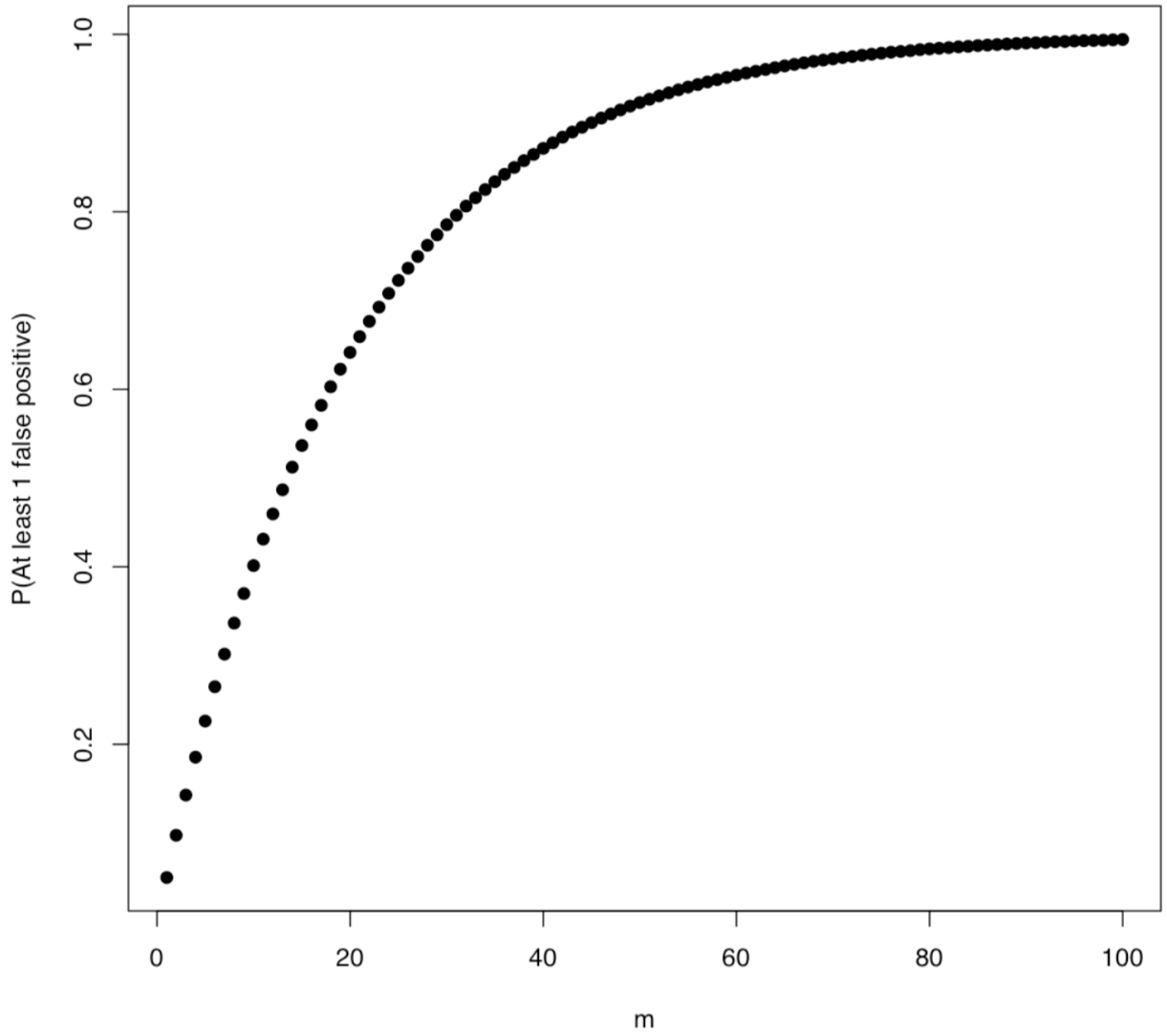
Multiple testing correction

- Because you are testing many hypotheses, you are likely to find statistically significant result simply by chance
- If you are testing for differential expression for all genes, how many are differentially expressed under the null at α =0.05?

“Controlling Type I error”

- If there are *m* null hypothesis tests, what is the probability of at least 1 false positives, assuming that P(making an error) = α

Probability of making at least 1 FP call



Bonferroni correction

- One option is to control this Family-Wise Error Rate (the probability of at least one type I error)
- Set the significance cut-off at α/n.
- Probability of making at least one error now?

- It is too conservative (too many false negatives)
- Counter-intuitive: interpretation of finding depends on the number of tests?
- Do you want to guard against ANY false positives?
- The general null hypothesis (all null hypotheses are true) is rarely of interest.

False Discovery Rate

- In many large-scale experiments, **we can tolerate some false positives**
- FWER is appropriate when you want to guard against ANY false positives
- Thus a popular alternative is control the false discovery rate (FDR)

Controlling the false discovery rate (FDR)

	H_0 is true	H_1 is true	Total
Reject H_0	**V**	**S**	R
Not reject H_0	**U**	**T**	m - R
	m_0	$m-m_0$	m

- **FDR = V/R**
- FDR: The expected rate of incorrectly rejected null hypotheses ("false discoveries")
- Less stringent than the "family-wise error rate" approaches
- FDR is designed to control the **proportion of false positives among the set of rejected hypotheses** rather than to control the Type I error rate

Benjamini-Hochberg FDR

- To control FDR at level q:
- Order the unadjusted p-values: $p_1 \leq p_2 \leq \ldots \leq p_m$ for hypotheses $H_1, H_2, \ldots, H_m$.
- Find the test with the highest rank k for which the p-value is less or equal to $(k/m) * q$
- Reject all H_i for $i \leq k$

- On the right: q=0.05; m=10

Rank (k)	p-value	$(k/m) * q$	Reject
1	0.003	0.005	R
2	0.008	0.010	R
3	0.012	0.015	R
4	0.021	0.020	0
5	0.070	0.025	0
6	0.123	0.030	0
7	0.250	0.035	0
8	0.673	0.040	0
9	0.812	0.045	0
10	0.890	0.050	0

FDR vs pFDR

- When the test statistics are independent, this procedure controls the FDR at the level q.

Technical Details:

- Actually FDR ≤ $q^* m_0/m$. We can try to estimate m_0/m
- Also true under positive and negative correlations
- For highly correlated data, this may be conservative; use more powerful FDR procedure by resampling

- Benjamini-Hochberg: FDR = E[V/R | R>0] P(R>0)
- Storey & Tibshirani: pFDR = E[V/R | R>0]
- P(R>0) ~ 1 in nearly all cases and so the two are very similar

q-value

- “q-value”: the FDR analogue of the p-value
- q-value is the minimum FDR that can be attained when calling that feature significant (i.e., **expected proportion of false positives incurred when calling that feature significant**)
- The estimated q-value is a function of the p-value for that test and the distribution of the entire set of p-values from the family of tests being considered (Storey and Tibshirani 2003)
- Example: In a microarray study for differential expression, if gene X has a q-value of 0.04 it means that 4% of genes that show p-values less than or equal to that of gene X are false positives
- These q-values are still **estimates**
- We are typically more lenient with q-value cut-offs, e.g., 0.2

Simulation example

```
source("https://bioconductor.org/biocLite.R")
biocLite("qvalue")
library(qvalue)
```

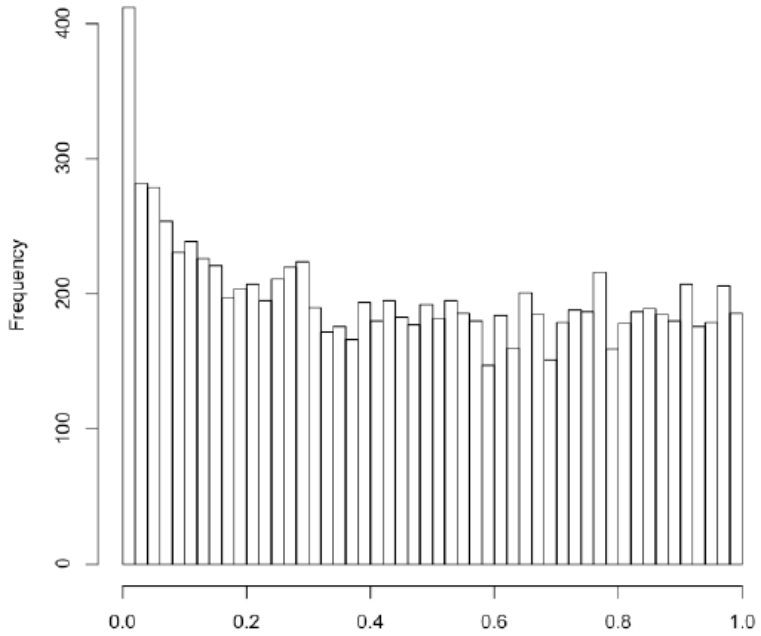
```
pv = NULL
n=10
for (i in 1:1000) {
  pv[i] = t.test(rnorm(n),rnorm(n))$p.value
}
for (i in 1001:1100) {
  pv[i] = t.test(rnorm(n),rnorm(n,mean=2))$p.value
}
h0 = 1:1000
h1 = 1001:1100

alpha = 0.05
e1 = sum(pv[h0]<alpha)/1000
e2 = sum(pv[h1]>alpha)/100
cat("Type I,II errors: ",e1,e2)

alpha1 = alpha/1100
e1 = sum(pv[h0]<alpha1)/1000
e2 = sum(pv[h1]>alpha1)/100
cat("Type I,II errors with Bonferroni: ",e1,e2)

qv <- qvalue(pv)$qvalues
e1 = sum(qv[h0]<alpha)/1000
e2 = sum(qv[h1]>alpha)/100
cat("Type I,II errors with FDR: ",e1,e2)
```

	H_0 is true	H_1 is true
Do not reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

Multiple Testing Correlations

- So what is the procedure in practice?
- Should the significance of my gene depend on that of other genes?

Plot the distribution of p-values



- What is the proper threshold for q-values?