Lecture 7: Logistic Regression and Survival Analysis

In this lecture we discuss when to use logistic regression and survival analysis, and learn how to perform these analyses in R.

Outline:
1. Why use logistic regression?
2. Overview of the logistic regression model.
3. How to perform logistic regression in R.
4. Why use survival analysis?
5. Overview of survival analysis models (Kaplan-Meier plots and Cox regression)
6. How to perform survival analysis in R.

Trivia:
What is the largest freshwater lake in the world (in volume)?
http://whc.unesco.org/en/list/754

By surface area?
http://geology.com/records/largest-lake.shtml

7.1 Logistic Regression
7.1.1 Why use logistic regression?

Last class we discussed how to determine the association between two categorical variables (odds ratio, risk ratio, chi-square/Fisher test). Suppose we want to explore a situation in which the dependent variable is dichotomous (1/0, yes/no, case/control) and the independent variable is continuous.

Let’s examine the Outbreak dataset in the epicalc library in R. Imagine we want to examine the association between vomiting (yes/no) as a dependent variable and age as a continuous predictor.

Now let’s see what happens when we try to examine this association using linear regression. The standard linear regression model would be:

\[ E[ \text{Vomiting} ] = \beta_0 + \beta_1 \text{age} \]
Y = Vomiting is 0/1, so the expected value of a 0/1 random variable is $p$, which here is the probability of vomiting as a function of age. Note that probabilities always fall between 0 and 1.

If we run the linear regression model in R, what happens?

```r
> out<-read.table("/Users/Avery/Desktop/classes/720 2015/classes/wk7/code/out.txt", sep="\t", header=T)
> attach(out)

> model.1 <- lm(vomiting~age)
> plot(vomiting~age,main="not linear!",col="purple")
> abline(model.1,lty=2)
```

Of course, R will give you estimates of coefficients. But are they meaningful?

Now observe the diagnostic plots we learned in class 5:
> plot(model.1, which=1:4)

Notice anything amiss?

Just about all the assumptions of linear regression are violated in these plots: the residuals are not normal; there is a discernable pattern in the scale-location plots indicating heteroscedasticity; and there are outliers and points with a high Cook’s Distance, indicating instability in the coefficient beta.

This type of thing happens because linear regression can only work well if the dependent variable is continuous. Here the predicted values still fall between 0 and 1, which is good, since we’re modeling a probability, but in many cases we can get an expected value of the probability of being a case that is above 1 or below 0, which is clearly nonsense.
7.1.2 Overview of Logistic Regression

When the assumptions of linear regression are violated, oftentimes you can transform the independent or dependent variables and the result is normality. In logistic regression the dependent variable is transformed using what is called the *logit* transformation:

\[
\text{logit (vomiting)} = \log \left( \frac{\# \text{ of people who vomited}}{\# \text{ of people who did not vomit}} \right)
\]

Then the new *logistic* regression model becomes:

\[
\log(\text{odds of vomiting}) = \beta_0 + \beta_1 \times \text{age}
\]

Covariates/predictors can be continuous or categorical, and the outcome is binary: 0/1. Since Y is either 0 or 1, expected value of Y for a set of covariates X is thought of as “*the probability that event Y occurs, given the covariates X*.” So we have that if \( p \) is the probability of vomiting, then

\[
\log \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1 \times \text{age}
\]

In linear regression, we were able to predict the outcome Y given new data by plugging in covariates from new data into the model. In the logistic model we can plug in values and get estimates for odds, odds ratios, or predicted probabilities (more on this later).

In standard linear regression, the coefficients are estimated based on the *least-squares* criterion. Here, the estimates for the coefficients are performed via *Maximum Likelihood Estimation* (MLE).

Due to a transformed outcome, there is a concomitant change in the interpretation of the regression coefficients:

- \( \beta_0 \) is the *log odds* of vomiting when age is equal to 0 (not very meaningful here)
- \( \beta_1 \) is the increase/decrease in the *log odds* for a one-unit increase in age
- \( e^{\beta_1} \) is the *odds ratio* comparing the increase/decrease in odds for those with a one-unit increase in age compared to the standard group (if you’re rusty on that number \( e \), see this: [http://people.bu.edu/aimcinto/number.e.pdf](http://people.bu.edu/aimcinto/number.e.pdf))
Some examples:

log(odds of vomiting for those aged 20) = \( \beta_0 + \beta_1 \times 20 \)
log(odds of vomiting for those aged 21) = \( \beta_0 + \beta_1 \times 21 \)

Since \( e^x \) is the inverse of the natural logarithm, \( \log(x) \), we have the following:

odds of vomiting for those aged 20 = \( e^{\beta_0 + \beta_1 \times 20} \)
odds of vomiting for those aged 21 = \( e^{\beta_0 + \beta_1 \times 21} \)

If we want to determine the odds ratio to compare the odds of vomiting for those who are 20 years old versus the odds of vomiting for those who are 21 years old we can do the following:

\[
\text{Odds ratio} = \frac{\text{odds of vomiting for those aged 21}}{\text{odds of vomiting for those aged 20}} = \frac{e^{\beta_0 + \beta_1 \times 21}}{e^{\beta_0 + \beta_1 \times 20}}
\]

By the laws of exponents (recall your precalculus), we can combine terms, and we get:

\[
e^{(\beta_0 + \beta_1 \times 21 - \beta_0 - \beta_1 \times 20)} = e^{\beta_1 (21 - 20)} = e^{\beta_1}
\]

Thus \( e^{\beta_1} = \exp(\beta_1) \) represents the odds ratio of vomiting comparing any two individuals who differ by a one-unit change in age.

If we want to get a predicted probability (p) of the outcome of interest, given our covariates, then we can do a little algebra, and we have that

\[
p = P(Y = 1 | X) = \frac{e^{(\beta_0 + \beta_1 \times x)}}{1 + e^{(\beta_0 + \beta_1 \times x)}}
\]

There is an large field of statistical modeling called generalized linear models, where the outcome variable undergoes some transformation to enable the model to take the form of a linear combination, i.e. \( f(E[Y]) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k \) for some function \( f \).
Logistic regression is just one such type of model; in this case, the function, often called the link, is

\[ f(y) = \log \left( \frac{y}{1 - y} \right) \].

In standard linear regression, the link function is just the identity: \( f(y) = y \).

There is Poisson regression (for count and rate data, log link), Gamma regression (continuous outcome skewed and strictly greater than 0), Multinomial regression (multiple categorical outcomes), and many more.

If you are interested in these topics, SPH offers
- BS 853: Generalized Linear Models (logistic regression is just one class)
- BS 820: Logistic Regression and Survival Analysis
- BS 852: Statistical Methods in Epidemiology (covers some logistic and survival, conditional matching)

And in the math department there is
- MA 575: Linear Models (great course, hard)
- MA 576: Generalized Linear Models

A final note on Logistic Regression:
Once we have our model, we may want to use it to predict future cases based on their covariates. This technique is called classification, and is used extensively in many fields. It is a first topic in a very popular field known as machine learning.


To do classification diagnostics, we can look at how well our prediction works by looking at false-positive and false-negative rates. We can run through the probabilities and say “person A is a case if their model-predicted \( p > \) threshold \( p^* \) and is not a case otherwise.” We can then count up all the false-positives and true positives for each threshold \( p^* \), and devise what is called an ROC curve (receiver operating characteristics curve). The ROC curve was originally developed from signal processing and information theory. In this way we can choose the best cut point for classification of future observations based on their covariates.
We will not cover this topic here, but if you’re interested, here is a brief overview of generating ROC curves in R: http://www.r-bloggers.com/roc-curves-and-classification/

7.1.3 Logistic Regression in R

To perform logistic regression in R, you need to use the `glm()` function. glm stands for “general linear model.” To run the above logistic regression model in R, we use the following command:

```r
> summary(glm(vomiting ~ age, family = binomial(link = logit)))
```

Call:
`glm(formula = vomiting ~ age, family = binomial(link = logit))`

Deviance Residuals:
```
       Min        1Q   Median       3Q      Max
-1.0671   -1.0174   -0.9365   1.3395   1.9196
```

Coefficients:
```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.14173    0.10621  -1.334    0.182
age           -0.01544    0.00396  -3.893 9.89e-05  ***
```

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1452.3  on 1093  degrees of freedom
Residual deviance: 1433.9  on 1092  degrees of freedom

AIC: 1437.9

Number of Fisher Scoring iterations: 4

--
Type `names(name of your model object)` to get all the output the model can produce for you.

Testing Significance of the model:
First thing’s first. We want to know if our model is any good. To test the significance for the overall model we use the following command:

```r
> 1-pchisq(1452.3-1433.9, 1093-1092)
[1] 1.79058e-05
```
This will give us a p-value.

It turns out that a likelihood comparison between the two models, null vs. the model of interest, when compared as a ratio, is a random variable (random in the betas), and follows a Chi-Square distribution.

The input to this test is:

- deviance of “null” model (model with only an intercept term) minus deviance of current model (can be thought of as “likelihood”)
- degrees of freedom of the null model minus df of current model

This is analogous to the global F test for the overall significance of the model that comes automatically when we run the \texttt{lm()} command. This is testing the null hypothesis that the model is no better (in terms of likelihood) than a model fit with only the intercept term, i.e. that all beta terms are 0.

**Interpreting the model:**
The fitted logistic model for these data is:

$$\log(\text{odds(vomiting)}) = -0.14 - 0.02 \times \text{age}$$

This means that for a one-unit increase in age there is a 0.02 decrease in the log odds of vomiting. This can be translated to an odds ratio $e^{-0.02} = 0.98$. People in an age group one unit higher than a reference group are estimated to have, on average, 0.98 times the odds of vomiting. How do we test the association between vomiting and age statistically?

**Hypotheses:**
- $H_0$: There is no association between vomiting and age (the odds ratio is equal to 1—equivalently, the beta term is equal to 0).
- $H_1$: There is an association between vomiting and age (the odds ratio is not equal to 1—equivalently, the beta term is not equal to 0).
When testing the null hypothesis that there is no association between vomiting and age we reject the null hypothesis at the 0.05 alpha level ($z = -3.89$, p-value = $9.89e-05$). On average, the odds of vomiting are 0.98 times that of identical subjects in an age group one unit smaller.

### 7.1.4 Multiple Logistic Regression in R

As with linear regression, we can also perform logistic regression with multiple predictors, this is called *multiple logistic regression*.

For example, if we want to examine the association between the vomiting and both age and gender, we would use the following logistic regression model:

$$\log (\text{odds of vomiting}) = \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{gender}$$

To estimate this model, we use the following syntax (very similar to linear regression):

```
> summary(glm(vomiting~age+sex, family=binomial(link=logit)))
```

```
Call:
  glm(formula = vomiting ~ age + sex, family = binomial(link = logit))

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-1.1043 -1.0302 -0.9006  1.3102  1.9103

Coefficients:  Estimate  Std. Error z value Pr(>|z|)
  (Intercept)  -0.32856  0.133013  -2.470  0.0135 *
  age         -0.016567 0.004002 -4.139 3.49e-05 ***
  sex          0.319911  0.134759   2.374  0.0176 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1452.3  on 1093 degrees of freedom
Residual deviance: 1428.2  on 1091 degrees of freedom
  AIC: 1434.2

Number of Fisher Scoring iterations: 4
Now to evaluate the model:

**Global (Omnibus) Hypothesis**
- $H_0$: There is no association between vomiting and age and gender
- $H_1$: There is an association between vomiting and age and gender

\[
> \text{pchisq}(1452.3-1428.2, 1093-1091)
\]

[1] 5.844556e-06

We reject the null hypothesis that there is no association between vomiting and age and gender ($\chi^2 = 24.1$, p-value=5.84e-06).

**Main Effects Hypotheses**
- $H_0$: There is no association between vomiting and age after adjusting for gender
- $H_1$: There is an association between vomiting and age after adjusting for gender

We reject the null hypothesis that there is no association between vomiting and age after adjusting for gender ($z=-4.14$, p-value = 3.49e-05). For a one unit increase in age the log odds of vomiting decrease by 0.02 after adjusting for gender. Those with a one-unit increase in age have 0.98 times the odds of vomiting after adjusting for gender.

- $H_0$: There is no association between vomiting and gender after adjusting for age.
- $H_1$: There is an association between vomiting and gender after adjusting for age.

We reject the null hypothesis that there is no association between vomiting and gender after adjusting for age ($z=2.37$, p-value = 0.0176). Males have a 0.32 increase in the log odds of vomiting in comparison to females after adjusting for age. That is, Males have 1.38 times the odds of vomiting in comparison to females after adjusting for age.

And just like in linear regression, if you want to do ANCOVA, or fit a model with categorical predictor variables, you can use the same syntax as described in lecture 5. Just type `factor(variable of interest)` as a predictor in your `glm()` command, and R will create $k - 1$ “dummy” variables for a k-category predictor.

**Quick aside on categorical predictors:**
- R automatically decides for you which factor level will be the “reference” category. If you want to change what your reference category is you can re-format the variable using
the `relevel()` function or the `level()` function to assign the category a specific reference. This works for all types of regression.

Finally, when we are looking at whether we should include a particular variable in our model (maybe it’s a potential confounder), we can include it based on the “10% rule,” where if our estimate of interest changes more than 10% when we include the new covariate in the model, then we include that new covariate in our final model.

**When we evaluate confounding in logistic regression, we compare the exponential of the betas, not the un-transformed betas themselves!**

We don’t really care if the betas change, only if the odds ratios change. What investigator will ask you “Tell me about the log odds…”?

There is quite a lot more to this topic. You can investigate model selection (using AIC), conditional logistic regression for matched design studies, sensitivity analysis for classification, and much more. My own work focuses on extending the logistic model in a Bayesian framework.

There are other possibilities for dichotomous outcome data. Two competing models to the logistic model are the Probit model and the Complementary Log-Log model. Both of these links model a dichotomous outcome as a function of predictors. They each have their own advantages and drawbacks. For a nice overview of the three models, see: [http://data.princeton.edu/wws509/notes/c3.pdf](http://data.princeton.edu/wws509/notes/c3.pdf)

**Exercise 1. Test the hypothesis that being nauseated was not associated with sex and age (use a multiple logistic regression model). Test the overall hypothesis that there is no association between nausea and sex and age. Then test the individual main effects hypothesis (i.e. no association between sex and nausea after adjusting for age, and vice versa).**
7.2 Survival Analysis
7.2.1 Why use survival analysis?

Let’s examine the dataset *leukemia* in the “survival” package.

```r
> rm(list = ls())
> library(survival)
> data(leukemia)
> detach(Outbreak)
> data(package="survival",leukemia)
> attach(leukemia)
```

The dataset tracks survival in patients with Acute Myelogenous Leukemia. The question at the time was whether the standard course of chemotherapy should be extended (‘maintenance’) for additional cycles.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
<td>Survival or censoring time</td>
</tr>
<tr>
<td>status</td>
<td>censoring status (0 if an individual was censored, 1 otherwise)</td>
</tr>
<tr>
<td>x</td>
<td>was maintenance chemotherapy given? (“Maintained” if yes, “Not maintained” if no)</td>
</tr>
</tbody>
</table>

Suppose we want to examine the association between the length of survival of a patient (how long they survived leukemia) and whether or not chemotherapy was maintained. Let’s create a histogram of the survival times.

```r
> hist(time, xlab="Length of Survival Time", main="Histogram of Survival Time in AML Patients")
```

![Histogram of Survival Time in AML Patients](image-url)
The survival times are highly skewed due to the fact that there is a person who survived much longer than everyone else. In addition there were quite a few people who survived for fewer than 10 months. This type of skewed distribution is typical when dealing with survival data, and thus the normality assumption of linear regression is often violated, making linear regression inappropriate for these data.

In addition to skewness, we also observe what is known as censoring with survival data.

```r
> leukemia[1:5,

<table>
<thead>
<tr>
<th></th>
<th>time</th>
<th>status</th>
<th>x</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>1</td>
<td>Maintained</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>1</td>
<td>Maintained</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>0</td>
<td>Maintained</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>1</td>
<td>Maintained</td>
</tr>
<tr>
<td>5</td>
<td>23</td>
<td>1</td>
<td>Maintained</td>
</tr>
</tbody>
</table>
```

The third observation has a status of 0. This means that the person was followed for 13 months and after that was lost to follow up. So we only know that this patient survived AT LEAST 13 months, but we have no other information available about the patient’s status.

This type of censoring (also known as right censoring) makes linear regression an inappropriate way to analyze the data. It is really an even bigger problem than the normality violation.

Instead of using linear or logistic regression to model the time an even occurs, we instead use a new class of models to track the probability an event occurs after some time T.

### 7.2.2 Overview of Survival Analysis

One way to examine whether or not there is an association between chemotherapy maintenance and length of survival is to compare the survival distributions. This is done by comparing Kaplan-Meier plots.

```r
> survfit(Surv(time, status)~1)

Call: survfit(formula = Surv(time, status) ~ 1)

records n.max n.start events median 0.95LCL 0.95UCL
 23      23      23     18      27      18      45
```
This tells us that for the 23 people in the leukemia dataset, 18 people were uncensored (followed for the entire time, until occurrence of event) and among these 18 people there was a median survival time of 27 months (the median is used because of the skewed distribution of the data—the mean would not be very helpful).

The 95% confidence interval for the median survival time for the 18 uncensored individuals is (18, 45).

> `summary(survfit(Surv(time, status)~1))`

Call: `survfit(formula = Surv(time, status) ~ 1)`

<table>
<thead>
<tr>
<th>time</th>
<th>n.risk</th>
<th>n.event</th>
<th>survival</th>
<th>std.err</th>
<th>lower 95% CI</th>
<th>upper 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>23</td>
<td>2</td>
<td>0.9130</td>
<td>0.0588</td>
<td>0.8049</td>
<td>1.000</td>
</tr>
<tr>
<td>8</td>
<td>21</td>
<td>2</td>
<td>0.8261</td>
<td>0.0790</td>
<td>0.6848</td>
<td>0.996</td>
</tr>
<tr>
<td>9</td>
<td>19</td>
<td>1</td>
<td>0.7826</td>
<td>0.0860</td>
<td>0.6310</td>
<td>0.971</td>
</tr>
<tr>
<td>12</td>
<td>18</td>
<td>1</td>
<td>0.7391</td>
<td>0.0916</td>
<td>0.5798</td>
<td>0.942</td>
</tr>
<tr>
<td>13</td>
<td>17</td>
<td>1</td>
<td>0.6957</td>
<td>0.0959</td>
<td>0.5309</td>
<td>0.912</td>
</tr>
<tr>
<td>18</td>
<td>14</td>
<td>1</td>
<td>0.6460</td>
<td>0.1011</td>
<td>0.4753</td>
<td>0.878</td>
</tr>
<tr>
<td>23</td>
<td>13</td>
<td>1</td>
<td>0.5466</td>
<td>0.1073</td>
<td>0.3721</td>
<td>0.803</td>
</tr>
<tr>
<td>27</td>
<td>11</td>
<td>1</td>
<td>0.4969</td>
<td>0.1084</td>
<td>0.3240</td>
<td>0.762</td>
</tr>
<tr>
<td>30</td>
<td>9</td>
<td>1</td>
<td>0.4417</td>
<td>0.1095</td>
<td>0.2717</td>
<td>0.718</td>
</tr>
<tr>
<td>31</td>
<td>8</td>
<td>1</td>
<td>0.3865</td>
<td>0.1089</td>
<td>0.2225</td>
<td>0.671</td>
</tr>
<tr>
<td>33</td>
<td>7</td>
<td>1</td>
<td>0.3313</td>
<td>0.1064</td>
<td>0.1765</td>
<td>0.622</td>
</tr>
<tr>
<td>34</td>
<td>6</td>
<td>1</td>
<td>0.2761</td>
<td>0.1020</td>
<td>0.1338</td>
<td>0.569</td>
</tr>
<tr>
<td>43</td>
<td>5</td>
<td>1</td>
<td>0.2208</td>
<td>0.0954</td>
<td>0.0947</td>
<td>0.515</td>
</tr>
<tr>
<td>45</td>
<td>4</td>
<td>1</td>
<td>0.1656</td>
<td>0.0860</td>
<td>0.0598</td>
<td>0.458</td>
</tr>
<tr>
<td>48</td>
<td>2</td>
<td>1</td>
<td>0.0828</td>
<td>0.0727</td>
<td>0.0148</td>
<td>0.462</td>
</tr>
</tbody>
</table>

The following summary goes through each time point in the study in which an individual was lost to follow-up or died, and re-computes the total number of people still at risk (n.risk), the number of events at that time point (n.event), the proportion of individuals who survived up until that point (survival) and the standard error (std.err) and 95% confidence interval (lower 95% CI, upper 95% CI) for the proportion of individuals who survived at that point.

(Notice that between times 13 and 18 one event happened, but the number at risk went from 17 to 14. So of the 17 people, one died, and two were lost to follow-up.)

You can also plot the survival curves using the following commands:
> surv.aml <- survfit(Surv(time,status)~1)
> plot(surv.aml, main = "Plot of Survival Curve for AML Patients",
  xlab = "Length of Survival", ylab = "Proportion of Individuals who have Survived")

This plot shows the survival curve (also known as a Kaplan-Meier plot), the proportion of individual who have survived up until that particular time as a solid black line and the 95% confidence interval (the dashed lines). The crosses are for censored events.

Our original question was to examine the association between chemotherapy maintenance and length of survival. To do this task in the context of survival analysis we compare the survival curves of those who received chemotherapy maintenance and those who did not.

First, let’s examine how to compare the survival statistics and create Kaplan-Meier plots for each chemotherapy group.
> print(survfit(Surv(time,status)~x)) #before ~x was ~1

Call: survfit(formula = Surv(time, status) ~ x)

    records n.max n.start events median 0.95LCL 0.95UCL
x=Maintained  11   11      11     7      31     18     NA
x=Nonmaintained 12   12      12     11     23      8     NA

We see that there is a total of 11 people who were on maintained chemotherapy, 7 died, the median follow up was 31. The 95% confidence interval of survival time for those on maintained chemotherapy is (18, NA); NA in this case means infinity. A 95% upper confidence limit of NA/infinity is not uncommon in survival analysis.

> plot(survfit(Surv(time,status)~x), main = "Plot of Survival Curves by Chemotherapy Group", xlab = "Length of Survival",ylab="Proportion of Individuals who have Survived",col=c("blue","red"))

> legend("topright", legend=c("Maintained", "Nonmaintained"),fill=c("blue","red"),bty="n")
In order to determine if there is a statistically significant difference between the survival curves, we perform what is known as a **log-rank test**, which tests the following hypothesis:

$H_0$: There is no difference in the survival function between those who were on maintenance chemotherapy and those who weren’t on maintenance chemotherapy.

$H_1$: There is a difference in the survival function between those who were on maintenance chemotherapy and those who weren’t on maintenance chemotherapy.

The test is performed in R as follows (before we used `survfit`):

```r
> survdiff(Surv(time,status)~x)
```

Call:
```
survdiff(formula = Surv(time, status) ~ x)
```

```
N Observed Expected (O-E)^2/E (O-E)^2/V
x=Maintained 11 7 10.69 1.27 3.4
x=Nonmaintained 12 11 7.31 1.86 3.4
```

Chisq= 3.4 on 1 degrees of freedom, p= 0.0653

So when testing the null hypothesis that there is no difference in the survival function for those who were on chemotherapy maintenance versus those who were not on chemotherapy maintenance we **fail to reject** the null hypothesis $\chi^2_1 = 3.4$ with a p-value = 0.0653.

### 7.2.3 Proportional Hazard Models

Suppose we want to see if there is a difference in survival functions between two groups after adjusting for a potential confounder.

We will not go into much detail on this model today, but briefly, a proportional hazard model is given as follows:

We define the **hazard** of an event as the risk of that event, as the time frame shrinks to 0. Usually when we calculate risk ratios, we have some time in mind, either cross-sectional, or, say, risk of dying after a year for two groups. Hazard is the risk, taken as the time frame vanishes to time $t = 0$. 
\[ h(t|X) = h_0(t) e^{\beta_0 + \beta_1 x} \]

\( h_0(t) \) is the “baseline hazard,” which we don’t worry too much about, because when we look at the ratio of hazards for two conditions, we get the following:

**Hazard ratio** for individual with \( X = x \) vs. \( X = (x+1) \):

\[
\frac{h(t|X = x + 1)}{h(t|X = x)} = \frac{h_0(t) e^{\beta_0 + \beta_1(x+1)}}{h_0(t) e^{\beta_0 + \beta_1 x}} = e^{[\beta_0 + \beta_1(x+1)] - [\beta_0 + \beta_1 x]} = e^{\beta_1}
\]

This term is the hazard ratio for the event of interest for people with covariate \( x+1 \) vs. people with covariate \( x \). If the term is \( >1 \), then those people who have a one-unit increase in their covariate compared against a reference group are at a higher “risk” (hazard) for the event.

All this is just the basics of Cox PH models. There is a lot more to these models, including various assumptions that need to be tested for the model validity to hold, and issues in interpretation. However, for our purposes, just seeing how to run these models in R is enough.

Using a new dataset in the survival package called “cancer” we want to examine the survival in 228 patients with advanced lung cancer from the North Central Cancer Treatment Group. Performance scores rate how well the patient can perform usual daily activities.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>inst</td>
<td>Institution code</td>
</tr>
<tr>
<td>time</td>
<td>Survival time in days</td>
</tr>
<tr>
<td>status</td>
<td>censoring status, 1=censored, 2=dead</td>
</tr>
<tr>
<td>age</td>
<td>age in years</td>
</tr>
<tr>
<td>sex</td>
<td>1=Male, 2= Female</td>
</tr>
<tr>
<td>ph.ecog</td>
<td>ECOG performance score (0= good, 5=dead)</td>
</tr>
<tr>
<td>ph.karno</td>
<td>Karnofsky performance score (0=bad, 100=good) rated by physician</td>
</tr>
<tr>
<td>pat.karno</td>
<td>Karnofsky performance score as rated by patient</td>
</tr>
<tr>
<td>meal.cal</td>
<td>calories consumed at meals</td>
</tr>
<tr>
<td>wt.loss</td>
<td>Weight loss in last six months</td>
</tr>
</tbody>
</table>
First, let’s do some non-adjusted analysis:

**Exercise 2.** Test the null hypothesis that there is no difference in the survival function of patients with advanced lung cancer between males and females. Hint: use the `survdiff()` function.

Suppose we want to determine if there is an association between length of survival and sex after adjusting for age. The log-rank test discussed previously will only compare groups, it does not take into account adjusting for other covariates/confounding variables.

To adjust for other covariates we perform the Cox proportional hazards regression using the `coxph()` function in R. To perform exercise 2 using Cox regression we use the following commands:

```r
> detach(leukemia)
> attach(cancer)
> coxph(Surv(time,status)~sex)
```

Call:
`coxph(formula = Surv(time, status) ~ sex)`

```
             coef exp(coef) se(coef)    z      p
sex -0.531   0.588  0.167    -3.18 0.0015
```

Likelihood ratio test=10.6 on 1 df, p=0.00111 n= 228, number of events= 165

We see that when using the Cox regression to perform the test, the results are very similar to the log rank test ($\chi^2 = 10.6$ with p-value = 0.00111). The Cox regression estimates the hazard ratio of dying when comparing males to females. Here, sex is significantly related to survival (p-value = 0.00111), with better survival in females in comparison to males (hazard ratio of dying = exp(-0.531) = 0.588). Females have 0.588 times the hazard of dying in comparison to males.

To extend the cox regression to adjust for other covariates, we will extend this to test the following hypothesis:
H₀: There is no difference in the survival function when comparing males to females after adjusting for age.
H₁: There is a difference in the survival function when comparing males to females after adjusting for age.

```r
> coxph(formula = Surv(time, status) ~ sex + age)

Call:
coxph(formula = Surv(time, status) ~ sex + age)

                 coef exp(coef)  se(coef)     z      p
sex       -0.5131  0.59903   0.16746 -3.06 0.0022
age        0.0172  1.01721   0.00922  1.85 0.0650

Likelihood ratio test = 14.1  on 2 df, p=0.000857  n= 228, number of events= 165

When testing the null hypothesis that there is no difference in the survival function between males and females after adjusting for age we reject the null hypothesis ($z = -3.06$, $p$-value = 0.0022). After adjusting for age, females have significantly better survival in comparison to males. Females have 0.599 times the hazard of dying in comparison to males, adjusting for age (HR<1).

Again, this is only scratching the surface of this topic. If you want to know more, I suggest you take BS 820: Logistic Regression and Survival Analysis, taught by Professor Mike Lavalley.

If you want to read more on survival analysis in R, here is a nice brief tutorial I found online: [http://anson.ucdavis.edu/~hiwang/teaching/10fall/R_tutorial%201.pdf](http://anson.ucdavis.edu/~hiwang/teaching/10fall/R_tutorial%201.pdf)
Assignment:
Homework 7 due next Tuesday, Final Project report due next Wednesday

Things we did not cover (or only touched on):

- **Density Estimation** (i.e. what are we really sampling from?) This topic is covered in MA589: Computational Statistics.

- **Multiple Testing Procedures**: Methods to control Family-Wise Error rate, False Discovery rate; Sequential Testing: [http://people.bu.edu/aimcinto/seq.testing.pdf](http://people.bu.edu/aimcinto/seq.testing.pdf).

- **Bootstrapping, The Jackknife, Permutation methods** for CIs and bias estimation; Here’s a brief overview of *The Jackknife Method* I wrote for a class: [http://people.bu.edu/aimcinto/jackknife.pdf](http://people.bu.edu/aimcinto/jackknife.pdf)

- **Generalized Linear Models (GLMs)** and their various assumptions and interpretations. We touched only briefly on Logistic & Survival Models; there are many more GLMs. (BU offers BS853: GLMs, MA576: GLMs)

- **SO Much More on Linear Models**: this topic is vast. We could have covered nonparametric regression, more on assumptions and violations, residual analysis, generalized additive models, LOWESS, ridge regression, LASSO.

- **Models for correlated/structured data**: Mixed-Effects Models, GEE, repeated measures analysis. (BU offers BS857: Analysis of Correlated Data)

- **Power and sample size calculations** (loops are essential here)

- **R for Computer Scientists**: R in Unix; nuts and bolts of R as a computer language (cf. the book *Advanced R*: [http://adv-r.had.co.nz/](http://adv-r.had.co.nz/))

- **Even more graphical options**: *ggplot2* and *lattice* packages for even better graphics [http://www.statmethods.net/advgraphs/ggplot2.html](http://www.statmethods.net/advgraphs/ggplot2.html) [http://www.statmethods.net/advgraphs/trellis.html](http://www.statmethods.net/advgraphs/trellis.html)

- **Missing Data Methods**. (This topic is touched on in BS852: Statistical Methods for Epidemiology, BS861: Applied Statistics in Clinical Trials II, and BS857.)

- **Propensity Score methods** for non-randomized clinical trials, unbalanced observational studies. [https://cran.r-project.org/web/packages/nonrandom/vignettes/nonrandom.pdf](https://cran.r-project.org/web/packages/nonrandom/vignettes/nonrandom.pdf)
• Dates and times in R. These can be tricky to deal with. See here: http://www.stat.berkeley.edu/~s133/dates.html

• Bayesian Statistics: (See this article on the four schools of statistics for more on this topic: http://labstats.net/articles/overview.html). Profs. Sebastiani and Doros teach classes on this topic in the biostatistics department: BS845, BS855, BS830 (not strictly Bayesian, but a lot of Bayesian methods are used). There is also MA578 taught by the Prof. Luis Carvalho in the math department.

• Time Series data (The math department offers MA585: Time Series. The topic is also touched on in BS857.) Here is a free online book on R for time series: https://media.readthedocs.org/pdf/a-little-book-of-r-for-time-series/latest/a-little-book-of-r-for-time-series.pdf