1. (Access Data Analysis) Coming soon!

2. Even though coronary heart disease (CHD) remains the leading cause of death in Western countries and the evidence is substantial that males are at higher risk than females, the role of genetic factors versus the gender factor in CHD is still largely unknown. A study was performed to assess the gender risk of death from CHD, controlling for genetic factors. Towards this end, twins consisting of a male and a female were identified. The age at which a male twin died of CHD was recorded, as was the age at which the female twin died (of CHD). Some times were censored due to deaths from other causes. The data are as follows, where a + indicates a censored observation.

<table>
<thead>
<tr>
<th>Age of twin’s death from CHD (years)</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>63+</td>
<td></td>
</tr>
<tr>
<td>49+</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td>56+</td>
<td>70+</td>
<td></td>
</tr>
<tr>
<td>68</td>
<td>75</td>
<td></td>
</tr>
<tr>
<td>74+</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>69+</td>
<td>69+</td>
<td></td>
</tr>
<tr>
<td>70+</td>
<td>70+</td>
<td></td>
</tr>
<tr>
<td>67</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>74+</td>
<td>74+</td>
<td></td>
</tr>
<tr>
<td>81+</td>
<td>81+</td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td>75+</td>
<td>73+</td>
<td></td>
</tr>
</tbody>
</table>

(a) Use a matched-pair test to assess whether the hazards for the males and females are equal. If so, which gender has the greater hazard (i.e. excess deaths)?

Solution: The null hypothesis is that the hazards are equal between two genders, adjusting for twin-pair differences, and the alternative for which the test is most powerful is that there is a common proportional hazards relationship between males and females across the strata:

\[ H_0 : \lambda_{i1}(t) = \lambda_{i2}(t), \text{ for all } i \text{ and } t > 0 \ vs. \]
\[ H_A : \lambda_{i1}(t) = \phi \lambda_{i2}(t), \text{ with } \phi \neq 0 \text{ for all strata } i \text{ and } t > 0 \]

** Note: This test will pick up individual strata differences if they are strong enough, but the above alternative is the form that the test will have most power for detecting.

```r
> obs.time <- c(50,49,56,68,74,69,70,67,74,81,61,75,63,52,70,75,72,69,70,74,81,58,73)
> delta <- c(1,0,0,1,0,0,0,1,0,0,1,0,1,0,1,1,0,0,0,1,0,0,1,0,1,0)
> male <- rep(1:0,each=12)
> pairid <- rep(1:12,2)
> survdiff( Surv(obs.time,delta) ~ male + strata(pairid) )

N Observed Expected (O-E)^2/E (O-E)^2/V
male=0 12 5 5.5 0.0455 0.2
```
The test has a p-value of 0.6547 so the test is not significant. We do not reject the null hypothesis that the hazard functions are equal. We also notice that males had 0.5 excess of observed vs. expected deaths.

(b) For what factors are you adjusting in this test?

Solution: In this study we are adjusting for all variables that are alike within the pairs, such as prenatal treatments, environment and socioeconomic factors (provided the twins were raised together) and most important, the fact that siblings share half of their genetic material. This provides at least a partial adjustment for genetics.

3. Consider a study in which we are interested in comparing the distributions of time to some event (e.g., time from entry into the study until death) across subpopulations defined by known predictor variables \( \mathbf{x}_i \), \( i = 1, \ldots, n \), \( j = 1, \ldots, p \). Let the random variable \( T_i \) denote the true failure time of the \( i \)th individual and \( C_i \) be the maximum time that the \( i \)th individual can be observed (i.e., the censoring time for the \( i \)th individual). We will assume the probability distribution for \( C_i \) carries no information about the value of \( T_i \). Thus our data consists of observations on \( n \) independent subjects:

\[
Y_i = \min(T_i, C_i) = \text{the time the } i \text{th individual was observed}
\]

\[
\delta_i = I_{[Y_i = T_i]} = \text{an indicator of an observed failure}
\]

\[
= \begin{cases} 1 & \text{if } Y_i = T_i \text{ (failure observed)} \\ 0 & \text{else (censored observation)} \end{cases}
\]

Using the observed data (\( y_i, \delta_i, \mathbf{x}_i \)), suppose we wish to compare survival across subpopulations defined by \( \mathbf{X}_i \) using a Cox proportional hazards model of the form

\[
\lambda(t|\mathbf{x}_i) = \lambda_0(t) \exp\{\mathbf{\beta}^T \mathbf{x}_i\}.
\]

(a) In class, we showed that the log-partial likelihood and score vector were respectively given by

\[
\ell(\mathbf{\beta}) = \sum_{j=1}^{D} \mathbf{\beta}^T \tilde{x}_{(j)} - \log \left( \sum_{i \in \mathcal{R}_j} e^{\mathbf{\beta}^T \mathbf{x}_i} \right)
\]

\[
\tilde{U}_k(\mathbf{\beta}) = \sum_{j=1}^{D} x_{(j)k} - \bar{x}_{(j)k}, \quad k = 1, \ldots, p
\]

where \( D \) denotes the total number of observed death times and \( \bar{x}_{(j)k} = \sum_{i \in \mathcal{R}_j} x_{ik} w_{(j)k}(\mathbf{\beta}) \) with

\[
w_{(j)k}(\mathbf{\beta}) = \frac{e^{\mathbf{\beta}^T \tilde{x}_i}}{\sum_{i \in \mathcal{R}_j} e^{\mathbf{\beta}^T \mathbf{x}_i}}.
\]

Show that the partial likelihood observed information is given by

\[
\mathbf{I}^*_kh(\mathbf{\beta}) = \mathbf{I}_k^h(\mathbf{\beta}) = \sum_{j=1}^{D} \left\{ \sum_{i \in \mathcal{R}_j} (x_{ik} - \bar{x}_{(j)k})(x_{ih} - \bar{x}_{(j)h}) w_{(j)k}(\mathbf{\beta}) \right\}, \quad k, h = 1, \ldots, p.
\]
Solution:

\[
\mathbf{I}_{kk} (\tilde{\beta}) = - \frac{\partial \hat{U}_k (\tilde{\beta})}{\partial \beta_h}
\]

\[
= \sum_{j=1}^{D} \left( \sum_{i \in R_j} x_{ik} x_{ih} e^{\tilde{\beta}^T x_i} \right) \left( \sum_{i \in R_j} e^{\tilde{\beta}^T x_i} \right) - \left( \sum_{i \in R_j} x_{ik} e^{\tilde{\beta}^T x_i} \right) \left( \sum_{i \in R_j} x_{ih} e^{\tilde{\beta}^T x_i} \right)
\]

\[
= \sum_{j=1}^{D} \left\{ \sum_{i \in R_j} x_{ik} x_{ih} w_{ij}(\tilde{\beta}) - \left( \sum_{i \in R_j} x_{ik} w_{ij}(\tilde{\beta}) \right) \left( \sum_{i \in R_j} x_{ih} w_{ij}(\tilde{\beta}) \right) \right\}
\]

\[
= \sum_{j=1}^{D} \left\{ \sum_{i \in R_j} x_{ik} x_{ih} w_{ij}(\tilde{\beta}) - \tilde{x}_{(j)k} \tilde{x}_{(j)h} \right\}
\]

\[
= \sum_{j=1}^{D} \left\{ \sum_{i \in R_j} x_{ik} x_{ih} w_{ij}(\tilde{\beta}) - 2\tilde{x}_{(j)k} \tilde{x}_{(j)h} \right\}
\]

\[
= \sum_{j=1}^{D} \left\{ \sum_{i \in R_j} \left( x_{ik} x_{ih} - 2\tilde{x}_{(j)k} \tilde{x}_{(j)h} + \tilde{x}_{(j)k} \tilde{x}_{(j)h} \right) w_{ij}(\tilde{\beta}) \right\} \quad \text{(since } \sum_{i \in R_j} w_{ij}(\tilde{\beta}) = 1 \text{)}
\]

\[
= \sum_{j=1}^{D} \left\{ \sum_{i \in R_j} \left( x_{ik} x_{ih} - \tilde{x}_{(j)k} \right) \left( x_{ih} - \tilde{x}_{(j)h} \right) w_{ij}(\tilde{\beta}) \right\}
\]

(b) Write an R function `partLklhd()` which takes arguments

- `Y`: a list of length 2, whose first element, `Y$y`, is a vector of observation times $y_i$, and whose second element, `Y$delta`, is a vector of failure indicators $\delta_i$.
- `X`: a matrix in which the $i$th row contains $\bar{x}_i^T$.
- `beta`: the vector of regression parameters $\tilde{\beta}$.

and returns a list having named elements

- `lnlklhd`: the value of the log partial likelihood function $\ell(\tilde{\beta})$.
- `score`: the value of the score function $\hat{U}(\tilde{\beta})$.
- `inform`: the value of the partial likelihood observed information matrix $\mathbf{I}^*(\tilde{\beta})$.

Solution:

```r
partLklhd <- function( Y, X, beta ){
    y <- Y$y
    p <- length(beta)
    lnlklhd <- 0
    score <- rep(0,p)
    inform <- matrix(0, nrow=p, ncol=p)
    events <- which( Y$delta==1 )
    for( j in events ){
        lnlklhd <- lnlklhd + (X[j,]%*%beta - log(sum( exp(X[j,]*beta)*ifelse(y>y[j],1,0) )))
    }
}
```
\[
w.j \leftarrow \frac{\exp(X_j^T \beta) \cdot \text{ifelse}(y=y[j],1,0)}{\sum \exp(X_j^T \beta) \cdot \text{ifelse}(y=y[j],1,0)}
\]
\[
xbar.j \leftarrow \frac{t(X_j) \cdot w.j}{w.j}
\]
\[
\text{score} \leftarrow \text{score} + (X_j - xbar.j)
\]
\[
\text{inform} \leftarrow \text{inform} + (t(X)\%\%X \cdot \text{as.vector}(w.j)) - xbar.j \%\%t(xbar.j)
\]
\[
\text{return}\left(\text{list}(\text{lnlklhd}=\text{lnlklhd}, \text{score}=\text{score}, \text{inform}=\text{inform})\right)
\]

(c) Test your function using the laryngeal cancer data adjusting for age and stage of disease (as a factor) and compare your results with the \texttt{coxph()} function in R (use \texttt{names()} to see how to obtain the log partial likelihood, score, and information matrix.

\[
> \text{larynx} \leftarrow \text{read.table("http://www.ics.uci.edu/~dgillen/STAT255/Data/larynx.txt" )}
> \text{set.seed(12345)}
> \text{larynx$t2death} \leftarrow \text{larynx$t2death} + \text{runif( length(larynx$t2death), -.001, .001 )}
> \text{coxphfit} \leftarrow \text{coxph( Surv( t2death, death ) ~ age + factor(stage), data=larynx )}
> \text{coxphfit$loglik}$[1] -196.61 -187.37
> \text{coxphfit$score}$[1] 25.165
> \text{solve(coxphfit$var)}
\]
\[
\begin{array}{cccc}
[1,] & 5073.600 & -23.2401 & -14.5743 & 26.7339 \\
[2,] & -23.240 & 5.9962 & -2.3962 & -1.4452 \\
[4,] & 26.734 & -1.4452 & -3.3030 & 7.4783 \\
\end{array}
\]

## Results of partLklhd() under the null model (beta=0)
##
\[
> \text{null.rslt} \leftarrow \text{partLklhd( Y=list(y=larynx$t2death, delta=larynx$death), X=model.matrix(coxphfit), beta=rep(0,4) )}
\]
\[
> \text{null.rslt}
\]
\[
\text{$lnlklhd}$
\]
\[
\begin{array}{c}
[1,] -196.61
\end{array}
\]
\[
\text{$score$}
\]
\[
\begin{array}{c}
\text{age} & 113.9275 \\
\text{factor(stage)2} & -3.1039 \\
\text{factor(stage)3} & 2.9408 \\
\text{factor(stage)4} & 7.7376
\end{array}
\]
\[
\text{$inform$}
\]
\[
\begin{array}{cccc}
\text{age} & \text{factor(stage)2} & \text{factor(stage)3} & \text{factor(stage)4} \\
4971.9350 & -24.25807 & -4.12937 & 7.82053 \\
-24.2581 & 8.03875 & -2.82038 & -0.65806 \\
-4.1294 & -2.82038 & 10.05469 & -0.91367 \\
7.8205 & -0.65806 & -0.91367 & 2.93934
\end{array}
\]
### Check of score statistic

\[ t(\text{null.rslt}\$score)\%\%\text{solve(null.rslt}\$inform)\%\%\text{null.rslt}\$score \]

\[,1\]

\[ [1,] 25.165 \]

### Results of partLklhd() under the SPMLE

\[ \text{spmle.rslt} \leftarrow \text{partLklhd(Y=list(y=larynx}\$t2death, \text{delta=larynx}\$death), \text{X=model.matrix(coxphfit)}, \text{beta=coxphfit}\$\text{coef}) \]

\[ \text{spmle.rslt} \]

\$lnlklhd

\[,1\]

\[ [1,] -187.37 \]

\$score

\[ \text{[,1]} \]

\begin{align*}
\text{age} & : 0.0000e+00 \\
\text{factor(stage)2} & : 1.6931e-15 \\
\text{factor(stage)3} & : 3.3307e-16 \\
\text{factor(stage)4} & : 2.2204e-16
\end{align*}

\$inform

\begin{array}{cccc}
\text{age} & \text{factor(stage)2} & \text{factor(stage)3} & \text{factor(stage)4} \\
5073.600 & -23.2401 & -14.5743 & 26.7339 \\
-23.240 & 5.9962 & -2.3962 & -1.4452 \\
-14.574 & -2.3962 & 10.9793 & -3.3030 \\
26.734 & -1.4452 & -3.3030 & 7.4783
\end{array}

4. For this part of the assignment we will implement a Newton-Raphson algorithm to solve for the coefficients of Cox’s proportional hazards regression model that we considered above.

(a) Write an R function `newtraph()`.

**Solution:**

```r
newtraph <- function( Y, X, beta, liklhd, verbose=F, eps=1e-8, iter.max=200 ){
  beta.start <- beta
  lik.start <- lik.old <- liklhd( Y, X, beta.start )
  delta <- solve( lik.old$inform ) %*% lik.old$score
  iter <- 0
  while( (sum( abs(delta) < eps ) < length(beta)) & (iter <= iter.max) ){ 
    delta <- solve( lik.old$inform ) %*% lik.old$score
    lik.new <- liklhd( Y, X, beta=(beta + delta) )
    if( lik.new$lnlkld < lik.old$lnlkld ){ 
      delta <- delta/2
      beta <- beta - delta
      lik.old <- liklhd( Y, X, beta )
    } else{
      beta <- (beta + delta)
    }
  }
}
```
lik.old <- liklhd( Y, X, beta )
iter <- iter + 1
}
if( verbose==T ) cat( "iteration : ", iter, " ; beta = ", round( beta, 7 ), "\n" )
}
if( iter > iter.max ) cat( "Convergence not met before maximum iterations reached \n" )
rslt <- list( beta0=beta.start, lnlklhd0=lik.start,
converge=(iter <= iter.max), nbriter=iter, beta=beta,
lnlkhd = lik.old$lnlkhd, score=lik.old$score, inform=lik.old$inform )
return( rslt )
}

(b) Consider the prostate cancer data on the course webpage (Note: A complete description of the
data is given on the Data portion of the course webpage). Use your Newton-Raphson function
to fit a proportional hazards regression model of time to relapse as a function of log(nadir PSA),
bone scan score, and performance status in the data set. Starting estimates can be taken from
a linear regression of the logarithm of time in remission versus the predictors. The negative of
the coefficients can be used as $\hat{\beta}$. Provide an interpretation of your estimated parameters and
compare your results to those given by the R function coxph().

Solution:
> keep <- rowSums(is.na(psa[,c("obstime","inrem","bss","ps","nadirpsa")]))==0
> psa <- psa[keep,]
##
##### coxph() fit
##
> coxphfit <- coxph( Surv( obstime, inrem=="no" ) ~ I(log10(nadirpsa)) + factor(bss) + ps,
data=psa )

##
##### Obtain starting values
##
> beta0 <- -lm( log(psa$obstime) ~ model.matrix(coxphfit) )$coef
> beta0 <- beta0[2:length(beta0)]

##
##### newtraph() fit
##
> myfit <- newtraph( Y=list(y=psa$obstime, delta=psa$inrem=="no"),
+ X=model.matrix(coxphfit), beta=beta0, liklhd=partLklhd,
+ verbose=F, eps=1e-8, iter.max=20 )

##
##### Comparison of SPMLE’s
##
> coxphfit$coef
I(log10(nadirpsa))       factor(bss)2      factor(bss)3         ps
 0.958887          -0.041953          0.740657         -0.038121

> myfit
$beta0
model.matrix(coxphfit)I(log10(nadirpsa))  model.matrix(coxphfit)factor(bss)2

6
0.59339363   -0.50259929
model.matrix(coxphfit)factor(bss)3  
-0.00042717   -0.01915895

$lnlklhd0
$lnlklhd0$lnlklhd
 [1,] -97.429

$lnlklhd0$score
   [,1]
I(log10(nadirpsa))  9.8133
factor(bss)2 -1.5661
factor(bss)3  3.6986
ps -68.8621

$lnlklhd0$inform
       I(log10(nadirpsa)) factor(bss)2 factor(bss)3 ps
I(log10(nadirpsa))  22.9072 -0.6167  3.3244 -9.215
factor(bss)2 -0.6167  6.3331 -5.4459 -11.801
factor(bss)3  3.3244 -5.4459  7.3581 -14.952

$converge
[1] TRUE

$nbriter
[1] 5

$beta
   [,1]
I(log10(nadirpsa))  0.958887
factor(bss)2 -0.041953
factor(bss)3  0.740657
ps -0.038121

$lnlklhd
   [,1]
[1,] -94.19

$score
   [,1]
I(log10(nadirpsa)) 4.2273e-12
factor(bss)2 1.0094e-12
factor(bss)3 2.8126e-12
ps -3.8369e-11

$inform
     I(log10(nadirpsa)) factor(bss)2 factor(bss)3 ps
I(log10(nadirpsa))  22.6927 -1.4422  2.4931  38.2278
factor(bss)2 -1.4422  5.3960 -5.1402 -2.4341
factor(bss)3  2.4931 -5.1402  5.8456 -7.1846
ps  38.2278 -2.4341 -7.1846  3349.7582
As can be seen from the above output, the point estimates for the MPLEs match (as they should!). In terms of interpretations we have the following (note that in practice I would also be supplying confidence intervals with these interpretations, but did not ask for them on this problem):

- We estimate that the relative risk of death is approximately 2.6-fold higher ($e^{0.9589}$) comparing two groups of patients, one of which has a nadir PSA value 10 times that of the other but are otherwise similar with respect to their bone scan score and performance score.
- We estimate that the relative risk of death is approximately 4.1% lower for patients with a bone scan of 2, compared to patients with a bone scan of 1 that are similar with respect to nadir PSA and performance score.
- We estimate that the relative risk of death is approximately 2.1-fold higher for patients with a bone scan of 3, compared to patients with a bone scan of 1 that are similar with respect to nadir PSA and performance score.
- We estimate that the relative risk of death is approximately 3.7% lower for patients with a performance score one unit higher than another group of patients who are similar with respect to nadir PSA and bone scan score.

5. Again consider Cox’s proportional hazards regression model and suppose that we are interested in comparing the hazard function between two populations (ie. we are modeling a single binary covariate taking on the value 0 or 1 depending on group membership). Show that in the absence of ties, the score test resulting from the proportional hazards model is equivalent to the usual 2-sample logrank test.

Solution: The logrank test is given by:

$$T_{LR} = \left[ \sum_{j=1}^{D} (O_j - E_j) \right]^2 \sum_{k=1}^{K} V_j$$

$$= \left[ \sum_{j=1}^{D} (d_{1j} - y_{1j} y_{j}) \right]^2 \sum_{j=1}^{D} y_{j} (y_{j} - 1)\frac{y_{1j} y_{j}}{y_{j} (y_{j} - 1)}$$

and in the absence of ties (ie. $d_{1k} = I_{[x_j=1]}$ for the $j$-th death time and $d_j = 1$) this reduces to

$$T_{LR} = \left[ \sum_{j=1}^{D} (x(j) - y_{1j} y_{j}) \right]^2 \sum_{j=1}^{D} \frac{y_{1j} y_{j}}{y_{j} (y_{j} - 1)}$$

Now, the numerator of the score statistic when modeling a single binary covariate is given by

$$[\tilde{U}(0)]^2 \left[ \sum_{j=1}^{D} (x(j) - \sum_{i \in R_j} \frac{x_{ik}}{\sum_{i \in R_j} 1}) \right]^2$$

$$= \left[ \sum_{j=1}^{D} (x(j) - y_{1j} y_{j}) \right]^2$$

which is the same as the numerator of $T_{LR}$ in the absence of ties. Further, the denominator if the score statistic is given by
\[ \Gamma(0) = \sum_{j=1}^{D} \left\{ \sum_{i \in R_j} \left( x_i - \sum_{i \in R_j} x_i \sum_{i \in R_j} 1 \right)^2 \sum_{i \in R_j} 1 \right\} \]

\[ = \sum_{j=1}^{D} \left\{ \sum_{i \in R_j} \left( x_i - \frac{y_{1j}}{y_j} \right)^2 \frac{1}{y_j} \right\} \]

\[ = \sum_{j=1}^{D} \left\{ \sum_{i \in R_j} \left( x_i^2 - 2 \frac{x_i y_{1j}}{y_j} + \frac{y_{1j}^2}{y_j^2} \right) \frac{1}{y_j} \right\} \]

\[ = \sum_{j=1}^{D} \left\{ \sum_{i \in R_j} \left( x_i - 2 \frac{x_i y_{1j}}{y_j} + \frac{y_{1j}^2}{y_j^2} \right) \frac{1}{y_j} \right\} \]

\[ = \sum_{j=1}^{D} \left( y_{1j} - 2 \frac{y_{1j}^2}{y_j} + \frac{y_{1j}^2}{y_j^2} \right) \frac{1}{y_j} \]

\[ = \sum_{j=1}^{D} \frac{y_{1j}(y_j - y_{1j})}{y_j^2} = \sum_{j=1}^{D} \frac{y_{1j} y_{0j}}{y_j^2}. \]