4. Alternate Derivation:
   a. Let \( R(i) \) be uncensored and unfailed obsn at time \( t_i \)
      (including \( i \))
   b. \( P \{ \text{obsn } i \text{ fails}\} | R(i) \) at risk, fail at time \( z \in R(i) \)
   c. Multiply to get Partial likelihood.
   d. Profile likelihood is \( L(\beta) = \prod_{i \in C} \sum_{k \in R(i)} \exp(z_{ik}) \)
      i. Definition makes sense for unordered times if risk set
         is defined to be all those not failed at time \( t_i \)
         • Requires censoring distributions the same for all
           values of \( \beta \)
      CO: 7.2
   ii. Also is conditional likelihood if censoring happens
      instantaneously after failure
   e. Can show that properties of likelihood extend to partial
      likelihood
5. Partial likelihood loses information contained in censoring.
   a. Example: Two data sets, having the same partial
      likelihood.
      i. Two groups, red \((z = 1)\) and blue \((z = 0)\).
         • See Fig. 9.
      ii. Data identical except for consecutive censored items
         swapped.
      iii. In fully-parametric model, right panel is stronger evidence favoring red.
      iv. Partial likelihood treats them identically:

\[
\pi_{ik} = \frac{\exp(z_{ik})}{\sum_{m \in R(i)} \exp(z_{im})} = \exp(z_m) \left( \sum_{m \in R(i)} \exp(z_m) \right)^{-1}
\]
   i. Indices:
      • \( i \) represents individual with event, \( k \) represents
        an individual (including \( i \)) at risk when \( i \) has
        event,
      • and \( j \) will represent a component of the the
        covariate vector, \( m \) will represent another
        individual at risk
   ii. Then \( \frac{\partial}{\partial \beta} \pi_{ik} \) is

\[
-\exp(z_{ik}) \left[ \sum_{m \in R(i)} \exp(z_m) z_{mj} \right] \left[ \sum_{m \in R(i)} \exp(z_m) \right]^{-2} + \exp(z_{ik} z_{kj} \left[ \sum_{m \in R(i)} \exp(z_m) \right]^{-1} - \pi_{ik} (z_{ij} - \sum_{m \in R(i)} \pi_{jm} z_{mj})
\]
   iii. Vector form \( \pi_{ik} = \pi_{ik} (z_k^T - \sum_{m \in R(i)} \pi_{im} z_m^T) \)
   c. \( \ell(\beta) = \sum_{i \in C} z_{ij} - \log(\sum_{k \in R(i)} \exp(z_{ik})) \)
   d. \( \ell'(\beta) = \sum_{i \in C} \left[ z_i - \sum_{k \in R(i)} z_{ik} \pi_{ik} \right] / \sum_{k \in R(i)} \exp(z_{ik}) \)

\[
\sum_{i \in C} \left[ z_i - \frac{\sum_{k \in R(i)} z_{ik} \pi_{ik}}{\sum_{k \in R(i)} \exp(z_{ik})} \right]
\]
   e. \( \ell''(\beta) = \sum_{i \in C} \left[ \frac{\sum_{k \in R(i)} z_{ik} \pi_{ik}}{\sum_{k \in R(i)} \exp(z_{ik})} \right] \frac{\partial}{\partial \beta} \sum_{k \in R(i)} \pi_{ik} \pi_{ik} \pi_{ik} + \sum_{k \in R(i)} \pi_{ik} \pi_{ik} \pi_{ik} \)
   KM: 8.3pn1

7. Estimator satisfies \( \ell'(\beta) = 0 \)
   a. Solve for \( \beta \) by using Newton–Raphson method.
      i. Guess \( \beta^0 \)
         • Generally start at zero.
      ii. \( \beta = \beta^0 \)
      iii. Solution \( \hat{\beta} = \beta - \ell''(\beta^0)^{-1} \ell'(\beta^0) \)
      iv. Update using new guess \( \hat{\beta} \rightarrow \beta^0 \)
   v. Repeat as needed:
      • Stop when \( \ell'(\beta^0) \) is sufficiently small.
      • Stop when update \( \ell''(\beta^0)^{-1} \ell'(\beta^0) \) is sufficiently small.
   b. proc phreg in SAS, coxph in R

8. What to do about ties?
   a. exact method of Cox
      i. Argue via conditioning
      ii. If tied variables covariates \( z_i, z_j \)
      iii. Replace \( \exp(z_{ij}) / \sum_{k \in R(i)} \exp(z_{ik})\) \( \exp(z_{ij}) / \sum_{k \in R(j)} \exp(z_{jk}) \) by
         \( \exp((z_i + z_j) / \sum_{k \in R(i)} \exp(z_{ik} + z_{ij}) / \sum_{k \in R(j)} \exp(z_{jk} + z_{ij})) \)
iv. \( \mathcal{P}(\{i, j\}) \) is the set of all pairs of individuals at risk at common failure time of \( i \) and \( j \).

v. Could be lots of them if more than two are tied.

b. Breslow’s method

   i. Approximate denominator by usual sum to power of number tied
   
   ii. Unfortunately a bit too big
   
   c. Efron method

   i. Product of sum and some adjustments.
   
   d. Which one you do is not so important.

C. Testing

1. Sampling distribution of the score statistic.

   a. Let \( U(\beta) = \ell'(\beta) \) be the.
   
   b. When data are independent observations \( Y_j \),

   \[
   \ell'(\beta) = \frac{d}{d\beta} \sum_{i=1}^{n} \log(p_j(Y_j, \beta)) = \sum_{i=1}^{n} \frac{d}{d\beta} q_j(Y_j, \beta)
   \]

   i. for \( q_j(y, \beta) = \log(p_j(y, \beta)) \).

   ii. \( \mathbb{E} \left[ \frac{d}{d\beta} \log(p_j(Y_j, \beta)) \right] = \sum_y \left( \frac{d}{d\beta} p_j(y, \beta) \right) p_j(y, \beta) = \sum_y \frac{d}{d\beta} p_j(y, \beta) = \frac{d}{d\beta} \sum y p_j(y, \beta) = \frac{d}{d\beta} 1 = 0 \)

   iii. Implies \( E_{\hat{\beta}}[U] \), under the true distribution.

   iv. As \( n \) increases, central limit shows that \( U(\beta) \) approximately multivariate normal in identically-distributed case.

   c. More general CLTs imply multivariate normality outside of the non-identically-distributed case, under some conditions.


   i. As in the Product Limit estimator, contributions to score are uncorrelated.

   e. Differentiating again,

   \[
   0 = \mathbb{E} \left[ \frac{d^2 \log(p_j(Y_j, \beta))}{d\beta_i d\beta_j} \right] = \sum_y \frac{d^2}{d\beta_i d\beta_j} q_j(y, \beta) p_j(y, \beta) = \sum_y \frac{d^2}{d\beta_i d\beta_j} q_j(y, \beta) p_j(y, \beta) + q_j^{\prime\prime}(y, \beta) p_j(y, \beta)
   \]

   i. \( q_j^{\prime\prime}(y, \beta) = \frac{d}{d\beta} q_j(y, \beta) \),

   \[
   q_j^{\prime\prime}(y, \beta) = \frac{d^2}{d\beta_i d\beta_j} q_j(y, \beta) ,
   \]

   f. Independence implies \( \text{Var}[U] = -\mathbb{E}[\ell''(\beta)] \).

   g. Often estimate \( \mathbb{E}[\ell''(\beta)] \) by its observed value.

   h. Hence \( U \approx N(0, -\ell''(\beta)) \).

2. Sampling Distribution for the estimator

   a. \( \hat{\beta} \approx \beta - \ell''(\beta)^{-1}\ell'(\beta) \).

   i. Newton Raphson iterations starting at (unknown) true value.

   b. Estimate variance of \( \hat{\beta} \) by

   \[
   (\ell''(\beta)^{-1})(-\ell''(\beta))(-\ell''(\beta)^{-1}) = \ell''(\beta)^{-1}
   \]

   c. Hence \( \hat{\beta} \approx N(\beta, -\ell''(\beta)^{-1}) \)

3. Likelihood Ratio and Wald Test Statistics

   a. \( H_0 : \beta = \beta^0 \) vs \( H_A : \beta \neq \beta^0 \)

   b. \( (\hat{\beta} - \beta^0)^\top(-\ell''(\hat{\beta}))^{-1}(\hat{\beta} - \beta^0) \) is Wald test statistic

   i. \( \hat{\beta} \approx N(\beta, -\ell''(\hat{\beta})^{-1}) \)

   ii. \( \hat{\beta} \) approximately unbiased.

   c. \( 2 \times \max \ell \) is (log) likelihood ratio test statistic.

4. Application to \( K \) sample problem

   a. Let \( z_{ij} = 1 \) if item \( i \) is in group \( j \)

   b. \( \beta^0 = 0 \).

   c. \( U_j(\beta) = \sum_{i \in C} \left[ z_{ij} - \frac{\sum_{k \in R(i)} z_{kj} \exp(z_{kj})}{\sum_{k \in R(i)} \exp(z_{kj})} \right] \)

   d. \( U_j(\beta^0) = \sum_{i \in C} \left[ z_{ij} - \frac{\sum_{k \in R(i)} z_{kj}}{\sum_{k \in R(i)} 1} \right] \)

   e. Same as log rank statistic

   i. \( \sum_{i \in C} z_{ij} = \sum_i D_{ij} \)

   ii. \( \sum_{k \in R(i)} 1 = Y_1 \)

   iii. \( D_1 = 1 \)

   iv. \( \sum_{k \in R(i)} z_{kj} = Y_{ij} \)

   f. As with ANOVA via regression, don’t treat the levels as a single variable with ordered categories.

5. Only \( K - 1 \) parameters are identifiable.

   a. Suppose \( \beta_1, \beta_2, \ldots, \beta_K \) are such that \( h_0(t) \exp(\beta_k) \) is the hazard for group \( k \).

   b. Let \( h_0^*(t) = \exp(\beta_1)h_0(t) \)

   c. Let \( \beta_1^* = \beta_k - \beta_0 \).

   d. Then \( h_0(t) \) and \( \beta \) and \( h_0^*(t) \) and \( \beta^* \) give the same probabilities.

   e. Often resolved by fixing \( \beta_1 = 0 \).

   f. Could also be resolved by fixing \( \beta_K = 0 \).

   g. Group whose parameter is set to zero is called baseline.

   h. Equivalent to dropping column from design matrix.
6. Tests for some parameters and not others:
   a. Notation: \( \beta = (\psi, \phi) \), \( \psi \) of interest.
   i. \( \phi \) not of interest
   b. Likelihood Ratio: \( 2 \times \) difference in maximized \( \ell \).
      i. Maximize \( \ell \) over \( \beta \)
      ii. Maximize \( \ell \) with \( \psi = 0 \) and \( \phi \) unconstrained.
         • degrees of freedom is length of \( \psi \)
   c. Wald Test \( (\psi - \psi_0)^T [\ell'^{11}(\hat{\beta})]^{-1} (\psi - \psi_0) \)
      i. Let \( \hat{\beta} = (\psi, \phi) \) maximize \( \ell \) over \( \beta \)
      ii. \( \text{Var}[\hat{\beta}] \approx I(\hat{\beta}) = [-\ell''(\hat{\beta})]^{-1} \)
      iii. \( \text{Var}[\hat{\psi}] \approx \text{appropriate sub-matrix of } I^{11} \)
   d. Score Test \( \ell^1(0, \tilde{\eta})^T I^{11}(0, \tilde{\eta}) \ell^1(0, \tilde{\eta}) \)
      i. Let \( \tilde{\phi} \) maximize \( \ell \) over \( \phi \) with \( \psi = 0 \)
      ii. Let \( \ell^1 \) be components of \( \ell' \) corresponding to \( \psi \)
   e. All three have approximate distribution \( \sim \chi^2 \), degrees of freedom is number of components in \( \psi \).
   f. Most common application: Test one parameter at a time
      i. Hence you can only estimate \( (K - 1)(L - 1) \) main effects.
      j. Software generally puts estimates it cannot estimate either to
         i. Missing
         ii. 0
         iii. omitted.
      R Code \( \text{SAS Code} \)

D. Estimation
1. Estimation is via maximum likelihood
   a. Estimate is most easily interpreted after exponentiating
      i. Called risk ratio
   b. For indicator variable, gives ratio of hazards in two groups
   c. For continuous variable, gives ratios of hazards for people identical except for the covariate taking values 1 unit apart
   C. 3.4
2. Confidence intervals
   a. Get confidence intervals for \( \beta_j \) as \( \hat{\beta}_j \pm z_{\alpha/2} \text{SE} \left[ \hat{\beta}_j \right] \)
      i. for \( \text{SE} \left[ \hat{\beta}_j \right] \) the same as in Wald test
   b. CI for \( \exp(\beta_j) \) may be calculated directly or on log scale
      i. \( \exp(\hat{\beta}_j) \pm \text{SE} \left[ \exp(\beta_j) \right] \) \( z_{\alpha/2} \)
         • Using delta method,
         \( \text{SE} \left[ \exp(\beta_j) \right] = \exp(\hat{\beta}_j) \text{SE} \left[ \beta_j \right] \)
   i. In this case, \( \psi \) has only one component, and is scalar.
   ii. Generally do inference via statistic before squaring.
   iii. Generally do Wald test.
   a. Allow the effect of one variable to depend on the effect of another.
   b. For right now, take both of these variables to be factors.
   c. Consider two factors, with \( K \) and \( L \) levels.
   d. These divide the observations into \( K \times L \) groups
   e. Interactions by definition allow a different behavior in each of the \( K \times L \) groups
      i. In the proportional hazards context, allow different hazard ratio, but same baseline hazard.
   f. This problem behaves like a single factor having \( K \times L \) levels.
   g. Again, only \( K \times L - 1 \) are identifiable.
   h. Usually parameterized in a way that makes sense if interactions are deemed unnecessary.
   i. Parameterized in terms of main effects,
      i. defined as the effect when the other variable is at reference level.
      ii. Again, no estimate is made for main effect at reference level.
      iii. So there are \( K - 1 + L - 1 \) main effects estimated.
      iv. Cannot estimate an interaction when either variable is at reference level.
   ii. or \( \exp(\hat{\beta}_j \pm z_{\alpha/2} \text{SE} \left[ \hat{\beta}_j \right]) \)
      • This one is probably better, since it doesn’t run into end of range.
   KM: 8.3pn2
3. Infinite Estimates
   a. If values of single covariate are in same order as event times, then estimator of associated \( \beta \) is \( \pm \infty \)
      i. Each term of \( \ell' \) for covariate \( j \) is
         \( z_{ij} - \sum_{k \in \mathcal{R}(i)} z_{kj} \exp(\alpha_k \beta) \)
      ii. Second part is weighted average of \( z_{kj} \)
      iii. In order to make weighted average = \( z_{ij} \) all weight must be on \( z_{ij} \)
      iv. If \( z_{kj} \leq z_{ij} \) for \( k \in \mathcal{R}(i) \) then \( \beta_k = \infty \) works
   b. Algorithm can’t converge in standard sense.
   c. Diagnose from convergence behavior.
      i. Warning message says algorithm hasn’t converged.
      ii. Or Very large parameter estimates.
   d. Also can happen with linear combination of \( \beta \)
      i. Harder to see by looking at data R Code \( \text{SAS Code} \)