EMORY
ROLLINS SCHOOL OF PUBLIC HEALTH
Department of Biostatistics and Bioinformatics

BIOS 522: Survival Analysis Methods

Reading 6:

Interpreting the Cox model

This week, we will interpret hazard ratios for binary, categorical, and continuous covariates. We will extend the allowable shapes by transformations, splines, and interactions. We will define global and local hypothesis tests for the Cox model and construct confidence intervals for hazard ratios.

Part 1. Interpreting hazard ratios

Point estimation for the hazard ratio

Recall the form of the Cox model:

$$h_i(t) = h_0(t)\exp(\beta_1 X_{i1} + \cdots + \beta_k X_{ik})$$

Recall the form of the Cox model partial likelihood

$$L(\beta) = \prod_{i=1}^{m} \frac{\exp(\beta_1 X_{i1} + \cdots + \beta_k X_{ik})}{\sum_{j \in R(T_i)} \exp\left(\beta_1 X_{j1} + \cdots + \beta_k X_{jk}\right)}$$

The partial likelihood is the product over failure times $T_1, \dots, T_m$, where the numerator is the based on the covariates of the person who failed at time T_i, and the denominator is based on the covariates of the people in the risk set $R(T_i)$.

We treat the partial likelihood as a typical likelihood to obtain our maximum partial likelihood estimates $\hat{\beta}_1, \dots, \hat{\beta}_k$. Recall that these are as log hazard ratios.

We are rarely interested in reporting the log hazard ratio directly. In scientific studies, we instead prefer to report the hazard ratio. To calculate the hazard ratio, we exponentiate our coefficient, e.g. $\exp(\hat{\beta}_1)$.

Binary covariates

For binary covariates ($X = 1$ or 0), the estimated hazard ratio $\exp(\hat{\beta})$ characterizes how much higher the hazard is in the group with $X = 1$ than the reference group with $X = 0$, holding constant other covariates in the model. Under the proportional hazards assumption, this multiplicative increase in hazard is the same (constant) across all times t.

The hazard ratio is useful because it provides a single numerical summary of the magnitude of the difference between two groups. Kaplan-Meier estimation offers a visual summary of this difference, but the hazard ratio distills this to a single number (under the simplifying assumption of proportional hazards).

Categorical covariates

For a categorical covariate with k levels, we select a single level of the covariate as a reference and use $k - 1$ indicator/dummy variables to model the hazard ratio between each level and the reference.

$$h_i(t) = h_0(t)\exp(\beta_1 I[X_i = 1] + \cdots + \beta_{k-1} I[X_i = k - 1])$$

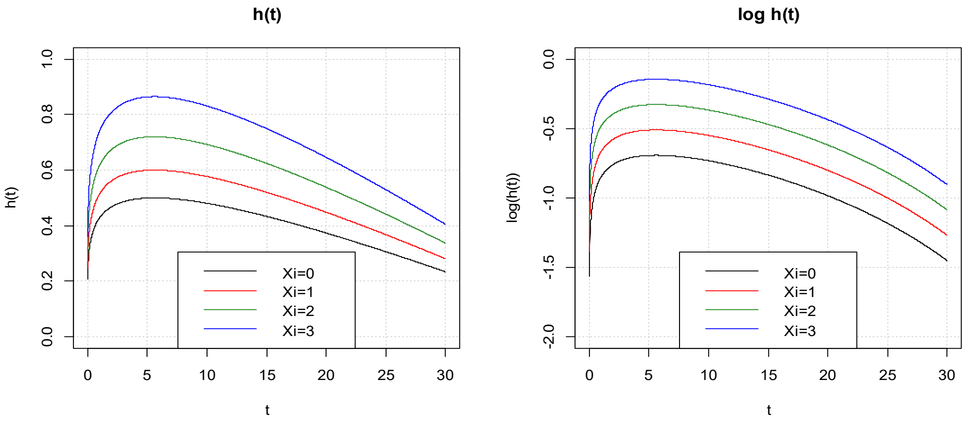
For **unordered categorical (nominal) variables**, the reference group can be selected arbitrarily, though commonly the largest group is selected.

For **ordered categorical (ordinal) variables**, typically the lowest (or highest) level of the variable is selected as the reference.

Continuous covariates

For continuous covariates, we interpret the coefficient as the log hazard ratio for a one-unit increase in the covariate. Like in linear or logistic regression, it is assumed that each one-unit increase in the covariate yields the same increase in the dependent variable. If the hazard doubles from ages 9 to 10 years, it doubles again from ages 10 to 11 years, and so on.

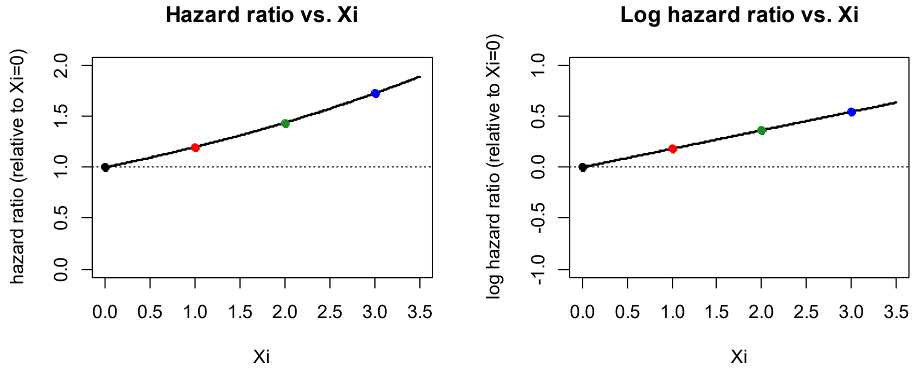
We can see this visually in the plots below. Consider a population with a single continuous covariate. Hazard functions (left) and log hazard functions (right) are plotted for individuals with $X_i = 0, 1, 2$ and 3.



In this example, the hazard ratio for X_i is 1.2. The hazard is multiplied by 1.2 with each one-unit increase in X_i. On the left-hand plot, note that each hazard function is 20% higher than the curve below it.

The log hazard ratio for X_i is 0.182. On the right-hand plot, note that the four log hazard functions are parallel. Each log hazard function is 0.182 units higher than the previous curve.

A logical way to summarize the effect of X_i is by plotting the hazard ratio or log hazard ratio (relative to the reference group) as functions of X_i.



In the left-hand plot, we see the hazard ratio (relative to the reference group with $X_i = 0$) for each value of X_i. The hazard ratio increases by 20% for each one-unit increase away from the reference group. This increases from 1.2 to $(1.2)^2$ =1.44 to $(1.2)^3$ =1.73. This increase is non-linear.

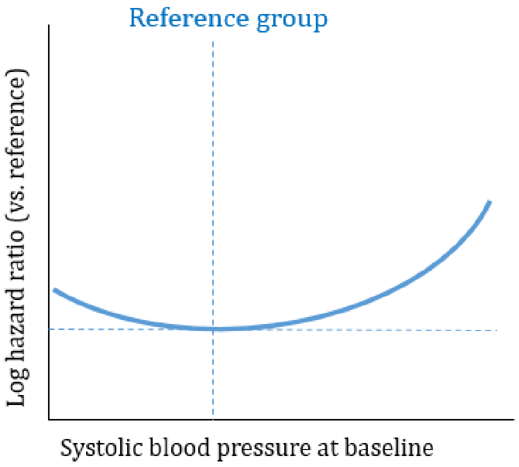
In the right-hand plot, we see the log hazard ratio (relative to the reference group with $X_i = 0$) for each value of X_i. The log hazard ratio increases at a constant slope of 0.182 for every one-unit increase away from the reference group. The log hazard ratio for the group with $X_i = 2$ is $2(0.182) = 0.364$. The log hazard ratio for the group with $X_i = 3$ is $3(0.182) = 0.547$. This increase is linear.

Thus, in Cox proportional hazards regression, the effect of the covariate X_i is modeled as a straight line on the log hazard ratio scale. This straight line crosses zero for the reference group, here $X_i = 0$.

Part 2. Interactions, splines, and transformations

Relaxing the linearity assumption

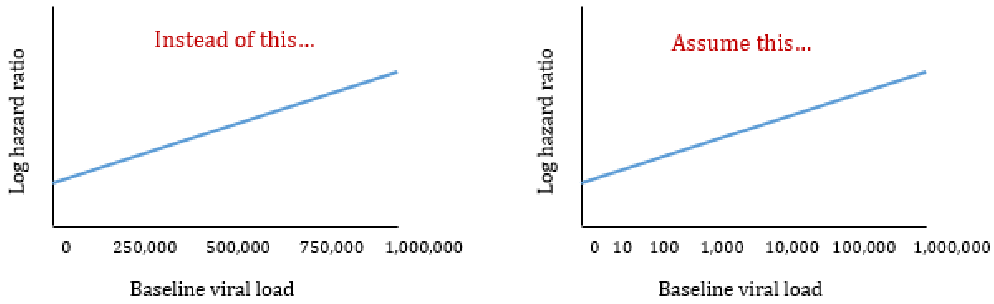
We can imagine situations where the relationship between a continuous covariate and the log hazard ratio is *not linear*. For example, it could have a U-shape. The picture below shows a hypothetical log hazard ratio for the relationship between time to stroke and baseline systolic blood pressure. Individuals with higher than normal systolic blood pressure are at elevated hazard of stroke. On the other hand, patients with extremely low systolic blood pressure are also at elevated hazard of stroke. The reference group is the group with normal systolic blood pressure (dotted vertical line). The log hazard ratio is zero at the dotted horizontal line since the hazard ratio for the reference group with itself is 1.



Transformations

In some settings, it may be necessary to **transform** a continuous covariate so that its effect is more linear on the log hazard ratio scale. Examples of common transformations include $\log(X_i)$, $\sqrt{X_i}$, and $\exp(X_i)$. The logic of transformations is same as described for other types of regression (linear, logistic). These transformations are appropriate when the relationship is monotonic (only increasing, or only decreasing) but is not adequately linear.

For example, measurements of viral load are often highly right skewed. HIV viral load can be >10,000,000 copies/ml, or it can be undetectable (<50 copies/ml). Where viral load is not broken into categories, it is common to take the log (base 10) of viral load before including it in a regression model. This more adequately captures the biological relationship between viral load and health.



Log transformation is also commonly applied to annual income. A \$10,000 increase in annual income is much more meaningful for someone making \$40,000 a year versus someone making \$1,000,000 a year. A log transformation can better capture the fact that, at higher income levels, only large increases have a large impact.

Polynomials and splines

We may want to flexibly model a continuous covariate to allow it to have a non-monotonic shape. Two common approaches include polynomials and splines.

The first strategy involves adding higher order terms (e.g. X_i^2, X_i^3) as separate covariates, allowing the relationship to take a **polynomial** shape.

Splines are even more flexible. We select multiple "knots" which define non-overlapping ranges of X_i values, and then we fit polynomials within each knot. **Penalized splines** are splines in which the complexity of the shape is constrained by a penalty term. Both achieve the same basic goal of greater flexibility in the relationship between the covariate and hazard.

Interactions

When we have more than one covariate, we frequently assume that the effects of the separate covariates are multiplicative (or additive on the log scale).

Imagine that we are modeling the hazard of stroke with two covariates, a binary covariate for treatment $X_{i,trt}$, and a continuous covariate for age $X_{i,age}$.

$$h_i(t) = h_0(t)\exp\left(\beta_{trt}X_{i,trt} + \beta_{age}X_{i,age}\right)$$

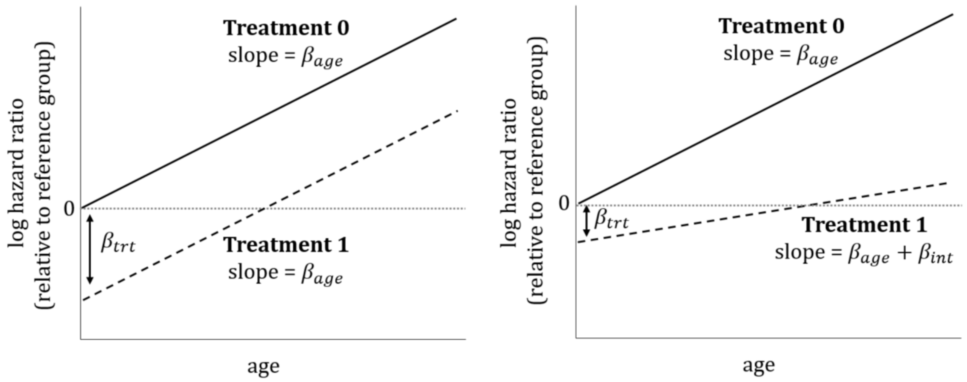
In the above model, we assume that the effect of treatment is the same for people of all ages. Regardless of age, the effect of treatment will be to multiply the hazard function by the same factor. Similarly, we assume that the effect of age is the same within each treatment group.

It may be possible, though, that the treatment is most protective in older individuals, and less so in younger individuals. We can model this by including an interaction term.

$$h_i(t) = h_0(t)\exp\left(\beta_{trt}X_{i,trt} + \beta_{age}X_{i,age} + \beta_{int}X_{i,trt}X_{i,age}\right)$$

The interaction term β_{int} allows the slope for age to vary for treated versus untreated populations.

The figure below summarizes the effect of age on the log hazard ratio for the two models. The left-hand figure does not include an interaction term. Note that the slope modeling the effect of age is the same in both treatment groups. The right-hand figure includes an interaction term. Though treatment always reduces the hazard, the difference between the treated and untreated groups is greatest for the oldest age group.



Interactions are commonly included between two binary covariates or between a binary covariate and a continuous covariate. Interactions can also be included between two continuous covariates in the form of a linear-by-linear interaction.

When including interactions in the model, it is important to also include terms for the underlying main effects. Failing to include these terms makes unusual assumptions. For example, failing to include the main effect for β_{age} in the example above assumes that that there is no effect of age (zero slope) in untreated individuals. The importance of retaining main effects in interaction models is known as the **hierarchical principle**.

Part 3. Hypothesis testing and interval estimation

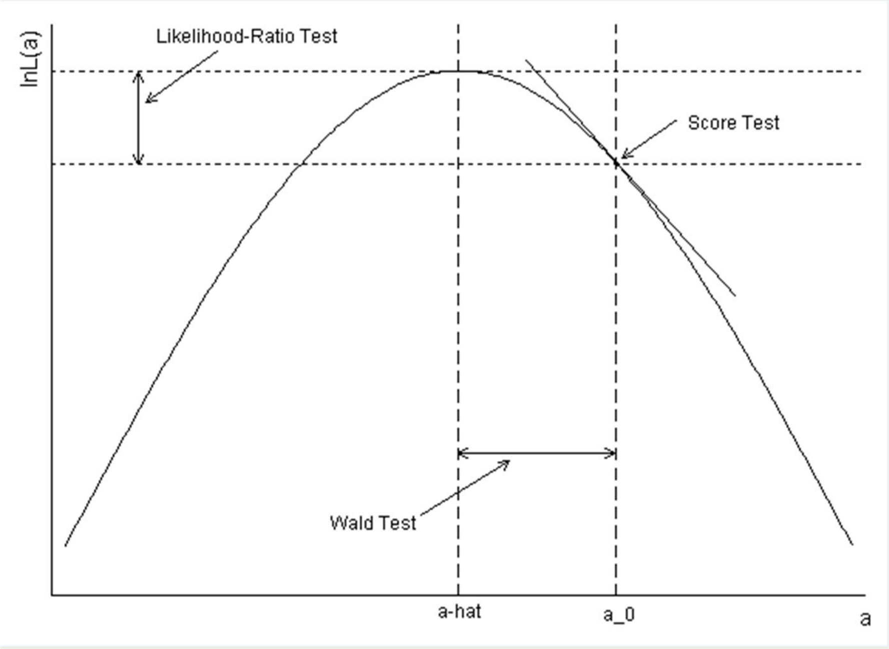
Hypothesis testing for the Cox model

There are several reasons why we would like to conduct hypothesis testing on our fitted Cox model. Some common examples are as follows:

- We may wish to test the null hypothesis that there is no association between a single covariate and survival, after adjusting for other covariates.
- We may wish to test the null hypothesis that there is no association between any within a group of covariates (e.g., dummy covariates representing a single categorical covariate) and survival, after adjusting for other covariates.
- We may wish to test the null hypothesis that a subset of groups has equal survival (e.g., to justify combining/collapsing groups within a categorical variable).
- We may wish to test for the presence of an interaction between covariates.

The above are examples of **local tests** – where the test is on a subset of coefficients within the model. For example, we may test the null that $\beta_1 = 0$ for covariate X_1, but place no restrictions on the coefficients for the remaining coefficients $X_2, \dots, X_p$. Or we might test that $\beta_1 = \beta_2 = \beta_3 = 0$ for the three dummy variables representing four educational levels in our data. Or we might test that $\beta_1 = \beta_2$ so that we can collapse the first two educational levels to simplify our model.

Local tests contrast with **global tests**. In a global test, we test that all p coefficients in our model are equal to a null value, like $\beta_1 = \beta_2 = \cdots = \beta_p = 0$.



There are three common *types* of tests we use for hypothesis testing in the Cox model. These are the Wald test, the likelihood ratio test, and the score test. A graphical representation is provided above for a model with a single coefficient a. The x-axis is a, with vertical lines denoting the null value a_0 and the maximum partial likelihood value $\hat{a}$. The y-axis is the log partial likelihood $\log L(a)$.

The **Wald test** assesses the difference between a_0 and $\hat{a}$. The **likelihood ratio test** compares the log partial likelihood evaluated at a_0 and $\hat{a}$, i.e., $\log L(a_0)$ and $\log L(\hat{a})$. The **score test** assesses whether the score function (slope of the partial likelihood function) is near zero, as it would be near the maximum likelihood value $\hat{a}$.

Local Wald test

A common local Wald test is to test the null hypothesis that a single covariate has no effect on survival after adjusting for all other covariates. We can construct a local Wald test for H_0: $\beta_j = 0$ from $\hat{\beta}_j$ and its standard error $\widehat{SE}(\hat{\beta}_j)$:

$$X_W = \frac{\hat{\beta}_j}{\widehat{SE}(\hat{\beta}_j)}$$

Under H_0, the test statistic X_W approximately follows a standard normal distribution. Details about the general form of the Wald test are provided in the Appendix.

Local likelihood ratio test

The likelihood ratio test can be used to compare nested models, with a common example being restricting one or a set of coefficients to equal 0. Consider two models: model 1 (M1) is our more general model, and model 0 (M0) is our restricted model, where some covariates are set to 0. We can calculate the log partial likelihood for each model, evaluated at its maximum partial likelihood estimate for each model. The likelihood ratio test has the following form:

$$X_{LR}^2 = 2\{\log L_{M1}(\hat{\beta}_{M1}) - \log L_{M0}(\hat{\beta}_{M0})\}$$

X_{LR}^2 approximately follows a χ_q^2 distribution under H_0, where q is the difference in the number of coefficients between the two models. Details about the general form of the likelihood ratio test are provided in the Appendix.

The score test – a special case!

The score test is based on the first derivative of the partial likelihood, also known as the score function $U(\cdot)$. Details about the general form of the score test are provided in the Appendix.

The score test for a Cox model has a special relationship with the log-rank test. In fact, in a particular setting, the two are equal!

Consider the special case where there is a single binary covariate $Z = 0$ or 1. When there are no ties, the score function for the Cox model is:

$$U(\beta) = \sum_{k=1}^{K} \left[Z_{(k)} - \frac{\sum_{j \in R(t_{(k)})} Z_j \exp(\beta Z_j)}{\sum_{j \in R(t_{(k)})} \exp(\beta Z_j)} \right]$$

To calculate the score test statistic, we evaluate the score function at the null value $\beta = 0$:

$$U(0) = \sum_{k=1}^{K} \left[Z_{(k)} - \frac{\sum_{j \in R(t_{(k)})} Z_j}{\sum_{j \in R(t_{(k)})} 1} \right]$$

It can be shown that the score test reduces to the log-rank test in the two-sample case. Define O_k as the number of subjects in group $Z = 1$ who fail at $t_{(k)}$. Define N_{1k} as the number of subjects in group $Z = 1$ at risk at $t_{(k)}$. Define N_k as the total number of subjects at risk at $t_{(k)}$. Since there are no ties, the number of failures at $t_{(k)}$ is $d_k = 1$. Though described for the setting with no ties, this holds whether or not ties are present.

Because there is a direct relationship between the Cox model and the log-rank test, it is common to see a log-rank test p-value reported in a table or figure alongside a hazard ratio from a Cox model.

Interval estimation for the hazard ratio

A common approach to calculate a confidence interval for the hazard ratio is to construct a Wald interval for the log hazard ratio and then exponentiate the endpoints. The confidence interval for β_j will take the form:

$$\left(\hat{\beta}_j - 1.96\,\widehat{SE}(\hat{\beta}_j), \hat{\beta}_j + 1.96\,\widehat{SE}(\hat{\beta}_j)\right)$$

To obtain a confidence interval for the hazard ratio, we exponentiate the endpoints. Thus, the confidence interval will take the form:

$$\left(\exp\left(\hat{\beta}_j - 1.96\,\widehat{SE}(\hat{\beta}_j)\right), \exp\left(\hat{\beta}_j + 1.96\,\widehat{SE}(\hat{\beta}_j)\right)\right)$$

Part 4. Looking ahead

After the midterm exam, we will continue our discussion of the Cox proportional hazards regression model. We will shift our focus to model fitting and diagnostics. We will discuss several different strategies for examining the proportional hazards assumption and overall goodness of fit.

Appendix: Hypothesis Testing (OPTIONAL READING)

This appendix provides formal definitions of several hypothesis tests. These are provided as reference. For ease of describing these tests, vector notation is used.

Consider fitting Cox model:

$$h(t|X) = h_0(t)\exp(\beta^T X)$$

where $\beta = (\beta_1, \dots, \beta_p)$ is a p-dimensional vector.

Global hypothesis testing

We could construct a **"global" hypothesis** of $H_0: \beta = \beta_0$ for some $\beta_0 = (\beta_{01}, \dots, \beta_{0p})$. For example, the hypothesis may be that all components are equal to zero. We refer to this as a "global" hypothesis because it involves all components of β.

We will consider three global hypothesis tests: a global Wald test, a global likelihood ratio test, and a global score test.

To construct these tests, the relevant quantities are:

- Partial likelihood: $L(\beta)$ (scalar)
- Log partial likelihood: $\ell(\beta) = \log L(\beta)$ (scalar)
- Score function: $U(\beta) = \frac{\partial \ell(\beta)}{\partial \beta}$ ($p \times 1$ vector)
- Information matrix: $I(\beta) = -\frac{\partial U(\beta)}{\partial \beta^T}$ ($p \times p$ matrix)
- Maximum partial likelihood estimate of β: $\hat{\beta}$:

The maximum partial likelihood estimator is:

$$\hat{\beta} = \text{argmax}_\beta L(\beta)$$

It can be obtained as the solution of $U(\beta) = 0$ via the Newton-Raphson algorithm. $\hat{\beta}$ is consistent and asymptomatically normal.

The first test we consider is a **global Wald test**. Given that:

$$\hat{\beta} \mathrel{\dot{\sim}} N(\beta, I(\beta)^{-1})$$

The global Wald test statistic is:

$$X_W^2 = (\hat{\beta} - \beta_0)^T I(\hat{\beta})(\hat{\beta} - \beta_0)$$

where $I(\hat{\beta})$ is the observed information. X_W^2 approximately follows a χ_p^2 distribution under H_0.

The next test is the **global likelihood ratio test**, which has test statistic:

$$X_{LR}^2 = 2\{\ell(\hat{\beta}) - \ell(\beta_0)\}$$

X_{LR}^2 approximately follows a χ_p^2 distribution under H_0.

The final test is the **global score test**, which has test statistic:

$$X_{SC}^2 = U(\beta_0)^T I^{-1}(\beta_0) U(\beta_0)$$

where $I^{-1}(\beta_0)$ is the *inverse* of the information matrix evaluated at the null value of β_0. Recall that the information is also the variance of the score function. X_{SC}^2 approximately follows a χ_p^2 distribution under H_0.

In general, the likelihood ratio test and Wald test perform similarly. The score test tends to inflate the size of the test.

Local tests of covariate effects

Alternatively, we may wish to locally test individual covariates or a subset of covariates. For example, for a categorical covariate with 4 levels, we may seek to test if the 3 coefficients added to model this covariate are all equal to zero.

Generally, one is interested in testing a hypothesis about a subset of $\beta = \{\beta_1, \dots, \beta_p\}$, say $\beta_J = \{\beta_j, j \in J\}$, where J is the index set of components of vector β of interest. β_J contains q components $(0 < q \leq p)$. Our local hypothesis is $H_0: \beta_J = \beta_{J0}$.

The first test we consider is a **local Wald test**, which has test statistic:

$$X_W^2 = \left(\hat{\beta}_J - \beta_{J0}\right)^T \left\{L^J\left(\hat{\beta}\right)\right\}^{-1} \left(\hat{\beta}_J - \beta_{J0}\right)$$

where $\hat{\beta}_J = (\hat{\beta}_j, j \in J)$ estimates the regression coefficients of interest, and $L^J(\beta)$ is the submatrix of $I^{-1}(\beta)$ that consists of $\left\{\left(I^{-1}(\beta)\right)_{jk}, j \in J, k \in J\right\}$. X_W^2 approximately follows a χ_q^2 distribution under H_0.

The next test is a **local likelihood ratio test**, which has test statistic:

$$X_{LR}^2 = 2\left\{\ell\left(\hat{\beta}\right) - \ell_{\beta_{J0}}\left(\hat{\beta}_{-J}\left(\beta_{J0}\right)\right)\right\}$$

Where $\ell_{\beta_{J0}}(\cdot)$ is the log partial likelihood derived when fixing $\beta_J = \beta_{J0}$, and $\hat{\beta}_{-J}\left(\beta_{J0}\right)$ is the restricted MPLE which maximizes $\ell_{\beta_{J0}}(\cdot)$. X_{LR}^2 approximately follows a χ_q^2 distribution under H_0.

The final test is the **local score test**, which has test statistic:

$$X_{SC}^2 = U_J\left\{\beta_{J0}, \hat{\beta}_{-J}\left(\beta_{J0}\right)\right\}^T I^J\left(\beta_{J0}, \hat{\beta}_{-J}\left(\beta_{J0}\right)\right) U(\beta_{J0}, \hat{\beta}_{-J}\left(\beta_{J0}\right))$$

Where $U_J(\cdot)$ is the subvector of $U(\cdot)$ that consists of $\left\{U_j(\cdot), j \in J\right\}$, and $I^J(\beta)$ is the submatrix of $I^{-1}(\beta)$ that consists of $\left\{\left(I^{-1}(\beta)\right)_{jk}, j \in J, k \in J\right\}$.
X_{SC}^2 approximately follows a χ_q^2 distribution under H_0.