8 Likelihood Ratio, Score, and Wald Tests

A good reference for this material is Section 9.3 of Cox and Hinkley. Chapter 7 of Silvey is also devoted to this topic, and some related material is in Section 6.6 of Bickel and Doksum and Section 8.4 of C&B.

Throughout this section assume that at the true parameter value both the full model and the restricted model given by the null hypothesis satisfy the conditions for consistency and asymptotic normality of the maximum likelihood estimates, see Section 6.

Suppose first that θ is 1-dimensional, and consider testing the simple null hypothesis $H_0 : \theta = \theta_0$. One way to proceed would be to calculate the MLE $\hat{\theta}_n$ and compare it to θ_0. Under the assumptions of Section 6, when the null hypothesis is true we have the large sample approximation

$$\hat{\theta}_n \approx \mathrm{n}(\theta_0, 1/I_n(\theta_0)),$$

where again "$\approx$" denotes "is approximately distributed", so we could use the statistic

$$(\hat{\theta}_n - \theta_0)[I_n(\theta_0)]^{1/2} \approx \mathrm{n}(0, 1)$$

to get an approximate test for the hypothesis. If we are only interested in testing for 2-sided alternatives, we could also use the statistic

$$W_n = (\hat{\theta}_n - \theta_0)^2 I_n(\theta_0) \approx \chi^2_1.$$

The test which rejects for large values of W_n is often referred to as the Wald test. Other consistent estimators of the asymptotic variance could also be used, and both $I_n(\hat{\theta}_n)$ or the observed information $-\partial^2 l(\hat{\theta}_n|\mathbf{X}_n)/\partial\theta^2$ can be used in place of $I_n(\theta_0)$. Both of these are consistent, as discussed in Section 6.3.

From Section 8.3.4 of C&B we know the locally most powerful test of $\theta = \theta_0$ is based on the score

$$\partial l(\theta_0|\mathbf{X}_n)/\partial\theta,$$

where $l(\theta|\mathbf{X}_n)$ is the log likelihood. Under the regularity conditions of Section 6 (of these notes), we know that when the true $\theta = \theta_0$, a large sample approximation to the distribution of the score is

$$\partial l(\theta_0|\mathbf{X}_n)/\partial\theta \approx \mathrm{n}(0, I_n(\theta_0)).$$

Thus we could base a test on the statistic

$$[I_n(\theta_0)]^{-1/2}\partial l(\theta_0|\mathbf{X}_n)/\partial\theta \approx \mathrm{n}(0, 1),$$

or the statistic

$$S_n = [I_n(\theta_0)]^{-1}[\partial l(\theta_0|\mathbf{X}_n)/\partial\theta]^2 \approx \chi^2_1,$$

where the distributions are under the null hypothesis. The test which rejects for large values of S_n is usually referred to as the score test. Again other consistent estimators of the asymptotic information could be used, such as the observed information (see Section 6.3).

From the development in Section 6.3, and in particular (6.8) and Theorem 6.4, we know that

$$\sqrt{n}(\hat{\theta}_n - \theta_0)[n^{-1}I_n(\theta_0)] - n^{-1/2}\partial l(\theta_0|\mathbf{X}_n)/\partial\theta \xrightarrow{P} 0$$

so

$$(\hat{\theta}_n - \theta_0)[I_n(\theta_0)]^{1/2} - [I_n(\theta_0)]^{-1/2}\partial l(\theta_0|\mathbf{X}_n)/\partial\theta \xrightarrow{P} 0,$$

at least under the null hypothesis. Because of this we might guess that the behavior of the tests should be similar in large samples. As discussed in Section 9, the difference in the two tests $\xrightarrow{P} 0$ under sequences of "local alternatives" as well, which means that the two tests are asymptotically equivalent, in the sense that asymptotically with probability 1, for a given outcome the two tests either both reject or both accept the null hypothesis.

A third test which could be used here is the large sample approximation to the likelihood ratio test. The test statistic is

$$Q_n = 2[l(\hat{\theta}_n|\mathbf{X}_n) - l(\theta_0|\mathbf{X}_n)].$$

To investigate the asymptotic distribution of this statistic, expand $l(\theta_0|\mathbf{X}_n)$ in a Taylor series about $\theta_0 = \hat{\theta}_n$. From Theorem 4.2, there is a value $\tilde{\theta}(\theta_0, \hat{\theta}_n)$ between θ_0 and $\hat{\theta}_n$ such that

$$l(\theta_0|\mathbf{X}_n) = l(\hat{\theta}_n|\mathbf{X}_n) + \frac{\partial l(\hat{\theta}_n|\mathbf{X}_n)}{\partial\theta}(\theta_0 - \hat{\theta}_n) + \frac{1}{2}\frac{\partial^2 l(\tilde{\theta}(\theta_0, \hat{\theta}_n)|\mathbf{X}_n)}{\partial\theta^2}(\theta_0 - \hat{\theta}_n)^2.$$

Noting that

$$\frac{\partial l(\hat{\theta}_n|\mathbf{X}_n)}{\partial\theta} = 0$$

(since the MLE satisfies the score equation) gives that

$$Q_n = -\frac{\partial^2 l(\tilde{\theta}(\theta_0, \hat{\theta}_n)|\mathbf{X}_n)}{\partial\theta^2}(\hat{\theta}_n - \theta_0)^2. \tag{8.1}$$

From Theorem 6.4, we know that

$$-n^{-1}\frac{\partial^2 l(\tilde{\theta}(\theta_0, \hat{\theta}_n)|\mathbf{X}_n)}{\partial\theta^2} \xrightarrow{P} \mathcal{I}(\theta_0) \tag{8.2}$$

when the null hypothesis is true, where $\mathcal{I}(\theta_0)$ is the limiting average information. Since

$$\mathcal{I}(\theta_0)[\sqrt{n}(\hat{\theta}_n - \theta_0)]^2 \xrightarrow{\mathcal{D}} \chi^2_1,$$

it follows from Slutsky's theorem that under the null

$$Q_n \xrightarrow{\mathcal{D}} \chi^2_1.$$

It also follows from (8.1) and (8.2) that

$$Q_n - W_n \xrightarrow{P} 0$$

when the null hypothesis is true. As with the score test this also holds under sequences of local alternatives, so that all 3 tests, Q_n, W_n, and S_n are asymptotically equivalent.

Example 8.1 Suppose $X_1, \ldots, X_n$ are a random sample from the exponential distribution with density $\exp(-x/\beta)/\beta$, and consider testing $H_0 : \beta = 1$ versus $H_1 : \beta \neq 1$. The log-likelihood is

$$l(\beta) = -n\log(\beta) - \sum_i X_i/\beta \tag{8.3}$$

and the score is

$$\frac{\partial l(\beta)}{\partial \beta} = -\frac{n}{\beta} + \frac{\sum_i X_i}{\beta^2}.$$

Setting this equal to 0 gives that the MLE is

$$\hat{\beta} = \overline{X}.$$

The observed information is

$$-\frac{\partial^2 l(\beta)}{\partial \beta^2} = -\frac{n}{\beta^2} + \frac{2\sum_i X_i}{\beta^3}.$$

Evaluating this at $\hat{\beta}$ gives

$$-\frac{\partial^2 l(\hat{\beta})}{\partial \beta^2} = -\frac{n}{\overline{X}^2} + \frac{2n}{\overline{X}^2} = \frac{n}{\overline{X}^2}.$$

The inverse of this (inverse of the observed information) is then a consistent estimator of the variance of $\hat{\beta}$. Noting that $\mathrm{E}(X_i) = \beta$, it follows that the expected information in the sample is

$$I_n(\beta) = \mathrm{E}\left(-\frac{\partial^2 l(\beta)}{\partial \beta^2}\right) = -\frac{n}{\beta^2} + \frac{2n\beta}{\beta^3} = \frac{n}{\beta^2}.$$

(As an aside, note that by direct calculation, $I_n(\beta)$ is also equal to $\mathrm{Var}[\partial l(\beta)/\partial\beta]$, since $\mathrm{Var}(X_i) = \beta^2$.) Thus

$$I_n(\hat{\beta}) = n/\overline{X}^2,$$

and $[I_n(\hat{\beta})]^{-1}$ also provides a consistent estimator of the asymptotic variance of the MLE. Although the observed and expected information are not the same, in this example when we evaluate them at the MLE they give the same estimator for the variance of the MLE. This will not always be true, but as noted in Example 6.1 it appears to be a general property of exponential families. Since the observed and expected information give the same variance estimator, it does not matter which we use in the Wald statistic. In either case we get the statistic

$$(\hat{\beta} - \beta_0)^2 I_n(\hat{\beta}) = (\overline{X} - 1)^2 (n/\overline{X}^2) = n\left(\frac{\overline{X} - 1}{\overline{X}}\right)^2 \approx \chi^2_1$$

under the null hypothesis.

For the score test it is conventional to evaluate the information at the null value $\beta = 1$. The reason for this is that you usually do a score test in part to avoid fitting the full model. From above,

$$I_n(1) = n/1^2 = n$$

and the observed information

$$-\frac{\partial^2 l(1)}{\partial\beta^2} = 2\sum_i X_i - n.$$

Note that these are not the same. Also note that it is possible to get data such that the observed information is negative (that is because it is evaluated at a fixed value of the parameter, which in general is not equal to the MLE—it will be positive when evaluated at the MLE). Because of this $-\partial^2 l(1)/\partial\beta^2$ may not be suitable as a variance estimator. The score itself is

$$\frac{\partial l(1)}{\partial\beta} = -n + \sum_i X_i,$$

so using $I_n(1)$ to estimate the variance, the score statistic is

$$\frac{\partial l(1)}{\partial\beta} \Big/ I_n(1) = n^{-1}\left(\sum_i X_i - n\right)^2 = n(\overline{X} - 1)^2 \approx \chi^2_1$$

The only difference between this and the Wald test is that the Wald test also has an $\overline{X}$ in the denominator. Since $\overline{X} \xrightarrow{P} \beta = 1$ under the null hypothesis, it follows that this difference is asymptotically negligible (under H_0).

For the likelihood ratio test, from (8.3),

$$2[l(\hat{\beta}) - l(1)] = 2[-n\log(\overline{X}) - n + n\overline{X}] = 2n[\overline{X} - 1 - \log(\overline{X})].$$

Again this should be approximately χ^2_1 under the null hypothesis. To check the asymptotic equivalence of this and the other two tests, expand $\log(x)$ in a second order Taylor series about $x = 1$ to get

$$\log(\overline{X}) \doteq (\overline{X} - 1) - (\overline{X} - 1)^2/2,$$

so

$$2[l(\hat{\beta}) - l(1)] \doteq n(\overline{X} - 1)^2,$$

which is the formula for the score test given earlier. Note that although the 3 statistics will all give the same results asymptotically, they generally give different values in finite samples. ◇

Exercise 8.1 Suppose $X_1, \ldots, X_n$ are iid $\mathrm{n}(0, \exp(2\gamma))$; that is, the density of X_i is

$$(2\pi)^{-1/2} e^{-\gamma} \exp(-x^2 e^{-2\gamma}/2).$$

Consider testing $H_0 : \gamma = 1$ versus $H_1 : \gamma \neq 1$.

(a). Find the MLE $\hat{\gamma}$.

(b). Give the 4 versions of the information: $I_n(\hat{\gamma})$, $I_n(\gamma_0)$, $-\partial^2 l(\hat{\gamma})/\partial\gamma^2$, and $-\partial^2 l(\gamma_0)/\partial\gamma^2$, where l is the log likelihood and $\gamma_0 = 1$ (the null value).

(c). Give the large sample Wald statistic using $I_n(\hat{\gamma})$, and specify the critical region for a test of approximate size α.

(d). Give the large sample score statistic using $I_n(\gamma_0)$, and specify the critical region for a test of approximate size α.

(e). Give the (large sample) likelihood ratio test statistic Q_n.

◇

The 1-dimensional case discussed above generalizes immediately to simple hypotheses with vector parameters. Assume $\boldsymbol{\theta}$ is k-dimensional, and treat all vectors as column vectors. All statements regarding distributions refer to the distribution under the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$. From Section 6.3 we know that subject to appropriate regularity conditions,

$$\hat{\boldsymbol{\theta}}_n \approx \mathrm{N}_k(\boldsymbol{\theta}_0, [\mathbf{I}_n(\boldsymbol{\theta}_0)]^{-1}),$$

so

$$W_n = (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)' \mathbf{I}_n(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \approx \chi^2_k,$$

see Theorem C.2. (The approximation in the distribution is that the distribution of the statistic converges to a χ^2_k. This can be shown formally using Theorem A.12, since W_n is a continuous function of $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$, which is asymptotically normal.) Again $\mathbf{I}_n(\boldsymbol{\theta}_0)$ in W_n could be replaced by $\mathbf{I}_n(\hat{\boldsymbol{\theta}}_n)$ or by the observed information matrix.

Similarly, if we set

$$[\mathbf{U}_n(\boldsymbol{\theta})]' = \left(\frac{\partial l(\boldsymbol{\theta}|\mathbf{X}_n)}{\partial \theta_1}, \ldots, \frac{\partial l(\boldsymbol{\theta}|\mathbf{X}_n)}{\partial \theta_k} \right),$$

then we also know that

$$\mathbf{U}_n(\boldsymbol{\theta}_0) \approx \mathrm{N}_k(\mathbf{0}, \mathbf{I}_n(\boldsymbol{\theta}_0)),$$

so

$$S_n = [\mathbf{U}_n(\boldsymbol{\theta}_0)]' [\mathbf{I}_n(\boldsymbol{\theta}_0)]^{-1} [\mathbf{U}_n(\boldsymbol{\theta}_0)] \approx \chi^2_k.$$

Also, from the expansion (6.9), we again have that

$$S_n - W_n \xrightarrow{P} 0.$$

The likelihood ratio statistic is again defined by

$$Q_n = 2[l(\hat{\boldsymbol{\theta}}_n|\mathbf{X}_n) - l(\boldsymbol{\theta}_0|\mathbf{X}_n)].$$

Using the multi-dimensional version of Taylor's theorem (Theorem 4.4), there exists a $\tilde{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_n)$ on the line segment joining $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_n$ such that

$$l(\boldsymbol{\theta}_0|\mathbf{X}_n) = l(\hat{\boldsymbol{\theta}}_n|\mathbf{X}_n) + [\mathbf{U}_n(\hat{\boldsymbol{\theta}}_n)]'(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_n) + \frac{1}{2}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_n)' \mathbf{H}_n(\tilde{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_n))(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_n),$$

where $\mathbf{H}_n(\boldsymbol{\theta})$ is the matrix of second partial derivatives with components

$$\frac{\partial^2 l(\boldsymbol{\theta}|\mathbf{X}_n)}{\partial \theta_i \partial \theta_j},$$

so

$$Q_n = (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)' [-\mathbf{H}_n(\tilde{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}_n))](\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$$

(note that again $\mathbf{U}_n(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}$). Using a generalization of the arguments in Theorem 6.4, it could then be shown that

$$Q_n - W_n \xrightarrow{P} 0,$$

and that

$$Q_n \approx \chi^2_k.$$

If we have a composite null hypothesis, the situation becomes more complex. Suppose we can partition $\boldsymbol{\theta}' = (\boldsymbol{\alpha}', \boldsymbol{\beta}')$, and the null hypothesis is $H_0 : \boldsymbol{\alpha} = \boldsymbol{\alpha}_0$, with the components of $\boldsymbol{\beta}$ being nuisance parameters. Suppose that $\boldsymbol{\alpha}$ has p components. Again only distributions under the null hypothesis will be considered, with $\boldsymbol{\beta}_0$ the true value of $\boldsymbol{\beta}$. It is convenient (almost essential) here to use vector and matrix notation, see Appendix B. All vectors are interpreted as column vectors. Since the asymptotic approximation

$$\begin{pmatrix} \hat{\boldsymbol{\alpha}}_n \\ \hat{\boldsymbol{\beta}}_n \end{pmatrix} \approx \mathrm{N}_k \left[\begin{pmatrix} \boldsymbol{\alpha}_0 \\ \boldsymbol{\beta}_0 \end{pmatrix}, [\mathbf{I}_n(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)]^{-1} \right], \tag{8.4}$$

is still valid, we can again develop a test based on the distribution of $\sqrt{n}(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0)$. To do this we need to extract the portion of $[\mathbf{I}_n(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)]^{-1}$ corresponding to $\boldsymbol{\alpha}$. Partition $\mathbf{I}_n(\boldsymbol{\alpha}, \boldsymbol{\beta})$ into blocks

$$\mathbf{I}_n(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \begin{pmatrix} \mathbf{I}_{n,\alpha\alpha}(\boldsymbol{\alpha}, \boldsymbol{\beta}) & \mathbf{I}_{n,\alpha\beta}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \\ \mathbf{I}_{n,\beta\alpha}(\boldsymbol{\alpha}, \boldsymbol{\beta}) & \mathbf{I}_{n,\beta\beta}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \end{pmatrix},$$

so that $\mathbf{I}_{n,\alpha\alpha}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is the covariance matrix of the $\boldsymbol{\alpha}$ components of the score, and so on. Note that $\mathbf{I}_{n,\alpha\beta}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = [\mathbf{I}_{n,\beta\alpha}(\boldsymbol{\alpha}, \boldsymbol{\beta})]'$. Using a formula for the inverse of a partitioned matrix (see Rao, *Linear Statistical Inference and Its Applications, Second Edition*, 1973, page 33), we can write the $\alpha\alpha$ block of $[\mathbf{I}_n(\boldsymbol{\alpha}, \boldsymbol{\beta})]^{-1}$ as

$$\{\mathbf{I}_{n,\alpha\alpha}(\boldsymbol{\alpha}, \boldsymbol{\beta}) - \mathbf{I}_{n,\alpha\beta}(\boldsymbol{\alpha}, \boldsymbol{\beta})[\mathbf{I}_{n,\beta\beta}(\boldsymbol{\alpha}, \boldsymbol{\beta})]^{-1}\mathbf{I}_{n,\beta\alpha}(\boldsymbol{\alpha}, \boldsymbol{\beta})\}^{-1}.$$

Define

$$\mathbf{I}_{n,\alpha\alpha|\beta}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathbf{I}_{n,\alpha\alpha}(\boldsymbol{\alpha}, \boldsymbol{\beta}) - \mathbf{I}_{n,\alpha\beta}(\boldsymbol{\alpha}, \boldsymbol{\beta})[\mathbf{I}_{n,\beta\beta}(\boldsymbol{\alpha}, \boldsymbol{\beta})]^{-1}\mathbf{I}_{n,\beta\alpha}(\boldsymbol{\alpha}, \boldsymbol{\beta}).$$

This is sometimes called the adjusted information (that is, the information for $\boldsymbol{\alpha}$ adjusted for having estimated $\boldsymbol{\beta}$). Then from (8.4), the marginal distribution of $\hat{\boldsymbol{\alpha}}_n$ is

$$\hat{\boldsymbol{\alpha}}_n \approx \mathrm{N}_p(\boldsymbol{\alpha}_0, [\mathbf{I}_{n,\alpha\alpha|\beta}(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)]^{-1}). \tag{8.5}$$

Since $\boldsymbol{\beta}_0$ is unknown, it will need to be estimated to get an estimate of the covariance matrix. Often both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ would be estimated by their MLEs, and the covariance matrix estimated by $[\mathbf{I}_{n,\alpha\alpha|\beta}(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n)]^{-1}$. The Wald test statistic then is

$$W_n = (\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0)'\mathbf{I}_{n,\alpha\alpha|\beta}(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n)(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0) \approx \chi^2_p.$$

Again any other consistent estimator of the asymptotic information could be used instead.

For the score statistic with nuisance parameters, partition the score vector $\mathbf{U}_n(\boldsymbol{\alpha},\boldsymbol{\beta})$ into $\mathbf{U}_{n,\alpha}(\boldsymbol{\alpha},\boldsymbol{\beta})$ containing the derivatives with respect to $\boldsymbol{\alpha}$ and $\mathbf{U}_{n,\beta}(\boldsymbol{\alpha},\boldsymbol{\beta})$ containing the derivatives with respect to $\boldsymbol{\beta}$. Since in general $\mathbf{U}_{n,\alpha}(\boldsymbol{\alpha},\boldsymbol{\beta})$ depends on the unknown nuisance parameters $\boldsymbol{\beta}$, it needs to be modified to be useful in hypothesis testing. One approach is to estimate $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}_0$, the MLE for $\boldsymbol{\beta}$ when $\boldsymbol{\alpha}$ is fixed at the null value $\boldsymbol{\alpha}_0$. That is, $\hat{\boldsymbol{\beta}}_0$ is found by maximizing

$$l(\boldsymbol{\alpha}_0,\boldsymbol{\beta}|\mathbf{X}_n)$$

over $\boldsymbol{\beta}$, or equivalently by solving

$$\mathbf{U}_{n,\beta}(\boldsymbol{\alpha}_0,\boldsymbol{\beta})=\mathbf{0}$$

for $\boldsymbol{\beta}$. Then we can try to use

$$\mathbf{U}_{n,\alpha}(\boldsymbol{\alpha}_0,\hat{\boldsymbol{\beta}}_0) \tag{8.6}$$

as the basis of a test statistic. To find the asymptotic approximation to the distribution of (8.6) we use informal methods. This can be made more rigorous using Taylor's formula and extensions of Lemma 6.1 and Theorem 6.4. First, taking a first order expansion and approximating the derivatives of the scores by their expected values we have

$$\begin{aligned}\mathbf{0} &= \begin{pmatrix}\mathbf{U}_{n,\alpha}(\hat{\boldsymbol{\alpha}}_n,\hat{\boldsymbol{\beta}}_n)\\ \mathbf{U}_{n,\beta}(\hat{\boldsymbol{\alpha}}_n,\hat{\boldsymbol{\beta}}_n)\end{pmatrix} \doteq \begin{pmatrix}\mathbf{U}_{n,\alpha}(\boldsymbol{\alpha}_0,\hat{\boldsymbol{\beta}}_0)\\ \mathbf{U}_{n,\beta}(\boldsymbol{\alpha}_0,\hat{\boldsymbol{\beta}}_0)\end{pmatrix} - \\ &\qquad \begin{pmatrix}\mathbf{I}_{n,\alpha\alpha}(\boldsymbol{\alpha}_0,\hat{\boldsymbol{\beta}}_0) & \mathbf{I}_{n,\alpha\beta}(\boldsymbol{\alpha}_0,\hat{\boldsymbol{\beta}}_0)\\ \mathbf{I}_{n,\beta\alpha}(\boldsymbol{\alpha}_0,\hat{\boldsymbol{\beta}}_0) & \mathbf{I}_{n,\beta\beta}(\boldsymbol{\alpha}_0,\hat{\boldsymbol{\beta}}_0)\end{pmatrix}\begin{pmatrix}\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0\\ \hat{\boldsymbol{\beta}}_n-\hat{\boldsymbol{\beta}}_0\end{pmatrix}.\end{aligned} \tag{8.7}$$

In the lower ($\boldsymbol{\beta}$) components of this expression, $\mathbf{U}_{n,\beta}(\boldsymbol{\alpha}_0,\hat{\boldsymbol{\beta}}_0)=0$, so all that is left is

$$\mathbf{I}_{n,\beta\beta}(\boldsymbol{\alpha}_0,\hat{\boldsymbol{\beta}}_0)(\hat{\boldsymbol{\beta}}_n-\hat{\boldsymbol{\beta}}_0)\doteq-\mathbf{I}_{n,\beta\alpha}(\boldsymbol{\alpha}_0,\hat{\boldsymbol{\beta}}_0)(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0),$$

or

$$(\hat{\boldsymbol{\beta}}_n-\hat{\boldsymbol{\beta}}_0)\doteq-[\mathbf{I}_{n,\beta\beta}(\boldsymbol{\alpha}_0,\hat{\boldsymbol{\beta}}_0)]^{-1}\mathbf{I}_{n,\beta\alpha}(\boldsymbol{\alpha}_0,\hat{\boldsymbol{\beta}}_0)(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0) \tag{8.8}$$

In the upper ($\boldsymbol{\alpha}$) components of (8.7) we have

$$\mathbf{U}_{n,\alpha}(\boldsymbol{\alpha}_0,\hat{\boldsymbol{\beta}}_0)\doteq\mathbf{I}_{n,\alpha\alpha}(\boldsymbol{\alpha}_0,\hat{\boldsymbol{\beta}}_0)(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)+\mathbf{I}_{n,\alpha\beta}(\boldsymbol{\alpha}_0,\hat{\boldsymbol{\beta}}_0)(\hat{\boldsymbol{\beta}}_n-\hat{\boldsymbol{\beta}}_0).$$

Substituting for $(\hat{\boldsymbol{\beta}}_n-\hat{\boldsymbol{\beta}}_0)$ from (8.8) then gives

$$\begin{aligned}\mathbf{U}_{n,\alpha}(\boldsymbol{\alpha}_0,\hat{\boldsymbol{\beta}}_0) &\doteq \mathbf{I}_{n,\alpha\alpha}(\boldsymbol{\alpha}_0,\hat{\boldsymbol{\beta}}_0)(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)\\ &\qquad -\mathbf{I}_{n,\alpha\beta}(\boldsymbol{\alpha}_0,\hat{\boldsymbol{\beta}}_0)[\mathbf{I}_{n,\beta\beta}(\boldsymbol{\alpha}_0,\hat{\boldsymbol{\beta}}_0)]^{-1}\mathbf{I}_{n,\beta\alpha}(\boldsymbol{\alpha}_0,\hat{\boldsymbol{\beta}}_0)(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)\\ &= \mathbf{I}_{n,\alpha\alpha|\beta}(\boldsymbol{\alpha}_0,\hat{\boldsymbol{\beta}}_0)(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)\\ &\doteq \mathbf{I}_{n,\alpha\alpha|\beta}(\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0)(\hat{\boldsymbol{\alpha}}_n-\boldsymbol{\alpha}_0)\end{aligned} \tag{8.9}$$

since $\hat{\boldsymbol{\beta}}_0 \xrightarrow{P} \boldsymbol{\beta}_0$. Using the asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0)$ from (8.5) then gives

$$\mathbf{U}_{n,\alpha}(\boldsymbol{\alpha}_0, \hat{\boldsymbol{\beta}}_0) \approx \mathrm{N}_p(\mathbf{0}, \mathbf{I}_{n,\alpha\alpha|\beta}(\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)).$$

Then we have

$$S_n = [\mathbf{U}_{n,\alpha}(\boldsymbol{\alpha}_0, \hat{\boldsymbol{\beta}}_0)]'[\mathbf{I}_{n,\alpha\alpha|\beta}(\boldsymbol{\alpha}_0, \hat{\boldsymbol{\beta}}_0)]^{-1}[\mathbf{U}_{n,\alpha}(\boldsymbol{\alpha}_0, \hat{\boldsymbol{\beta}}_0)] \approx \chi^2_p.$$

From (8.9), it also follows that

$$S_n - W_n \xrightarrow{P} 0.$$

Note that in S_n, the unknown parameters $\boldsymbol{\beta}_0$ in the information were estimated using the restricted estimate $\hat{\boldsymbol{\beta}}_0$. This is standard. The idea is that to perform the score test you would just fit the null model with $\boldsymbol{\alpha}$ fixed at $\boldsymbol{\alpha}_0$, while for the Wald test you would just fit the full model to compute the joint MLEs for both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

The likelihood ratio statistic for the composite null hypothesis is defined by

$$Q_n = 2[l(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n) - l(\boldsymbol{\alpha}_0, \hat{\boldsymbol{\beta}}_0)].$$

Note that this requires fitting both the null and alternative models. To approximate the distribution of Q_n in this setting, again expand in a Taylor series and approximate the second derivatives by their expected values to get

$$\begin{aligned} l(\boldsymbol{\alpha}_0, \hat{\boldsymbol{\beta}}_0) &\doteq l(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n) + [\mathbf{U}_n(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n)]' \begin{pmatrix} \boldsymbol{\alpha}_0 - \hat{\boldsymbol{\alpha}}_n \\ \hat{\boldsymbol{\beta}}_0 - \hat{\boldsymbol{\beta}}_n \end{pmatrix} - \\ &\quad \frac{1}{2} \begin{pmatrix} \hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0 \\ \hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_0 \end{pmatrix}' \begin{pmatrix} \mathbf{I}_{n,\alpha\alpha}(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n) & \mathbf{I}_{n,\alpha\beta}(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n) \\ \mathbf{I}_{n,\beta\alpha}(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n) & \mathbf{I}_{n,\beta\beta}(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n) \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0 \\ \hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_0 \end{pmatrix}, \end{aligned}$$

so

$$Q_n \doteq \begin{pmatrix} \hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0 \\ \hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_0 \end{pmatrix}' \begin{pmatrix} \mathbf{I}_{n,\alpha\alpha}(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n) & \mathbf{I}_{n,\alpha\beta}(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n) \\ \mathbf{I}_{n,\beta\alpha}(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n) & \mathbf{I}_{n,\beta\beta}(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n) \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0 \\ \hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_0 \end{pmatrix},$$

(since again $\mathbf{U}_n(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n) = \mathbf{0}$). Substituting (8.8) for $\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_0$, and multiplying out this expression (and simplifying), then gives

$$Q_n \doteq (\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0)' \mathbf{I}_{n,\alpha\alpha|\beta}(\hat{\boldsymbol{\alpha}}_n, \hat{\boldsymbol{\beta}}_n)(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0) = W_n.$$

Thus if we fill in the details, keep track of the remainder terms, and so on, we could show that

$$Q_n - W_n \xrightarrow{P} 0$$

and

$$Q_n \approx \chi^2_p.$$

Example 8.2 As an example of the composite hypothesis situation, consider the following exponential regression model. Suppose $Y_1, \ldots, Y_n$ are independent with densities

$$f_{Y_i}(y|\alpha, \beta) = \exp(\alpha + \beta z_i) \exp[-y \exp(\alpha + \beta z_i)],$$

where $\alpha, \beta \in R^1$ and the z_i are the known covariate values. Suppose we are interested in testing $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$ (that is, whether the covariate is associated with outcome).

The log-likelihood is

$$l = n\alpha + \beta \sum z_i - \sum y_i \exp(\alpha + \beta z_i),$$

the scores are

$$U_\alpha(\alpha, \beta) = \partial l/\partial\alpha = n - \sum y_i \exp(\alpha + \beta z_i)$$

and

$$U_\beta(\alpha, \beta) = \partial l/\partial\beta = \sum z_i[1 - y_i \exp(\alpha + \beta z_i)],$$

and the observed information terms are

$$-\partial^2 l/\partial\alpha^2 = \sum y_i \exp(\alpha + \beta z_i),$$

$$-\partial^2 l/\partial\alpha\partial\beta = \sum y_i z_i \exp(\alpha + \beta z_i),$$

and

$$-\partial^2 l/\partial\beta^2 = \sum y_i z_i^2 \exp(\alpha + \beta z_i).$$

Since $\mathrm{E}(Y_i) = 1/\exp(\alpha + \beta z_i)$, the expected information terms are particularly simple here:

$$I_{\alpha\alpha} = n, \quad I_{\alpha\beta} = \sum z_i, \quad I_{\beta\beta} = \sum z_i^2.$$

The information for β adjusted for α is

$$\begin{aligned} I_{\beta\beta|\alpha} &= I_{\beta\beta} - I_{\beta\alpha} I_{\alpha\alpha}^{-1} I_{\alpha\beta} \\ &= \sum z_i^2 - (\sum z_i)^2/n \\ &= \sum (z_i - \overline{z})^2. \end{aligned}$$

The MLEs would be found by setting the scores equal to 0 and solving for α and β. In general there will not be an explicit solution for the joint parameter vector (α, β), and iterative numerical methods would be needed to find the solution. If the joint MLEs are $(\hat{\alpha}, \hat{\beta})$, then the Wald statistic is just

$$\hat{\beta}' I_{\beta\beta|\alpha} \hat{\beta} = \hat{\beta}^2 \sum (z_i - \overline{z})^2$$

(recall β is a scalar).

We could avoid the need for iterative fitting by using the score test. When $\beta = 0$ the score equation for α is

$$n = \exp(\alpha) \sum y_i,$$

so

$$\hat{\alpha}_0 = \log[n/\sum y_i].$$

Then

$$U_\beta(\hat{\alpha}_0, 0) = \sum z_i[1 - y_i/\overline{y}],$$

and the variance of this score is approximately

$$I_{\beta\beta|\alpha} = \sum (z_i - \overline{z})^2.$$

The statistic

$$U_\beta(\hat{\alpha}_0, 0)/\sqrt{I_{\beta\beta|\alpha}} = \left(\sum z_i[1 - y_i/\overline{y}]\right) \left(\sum (z_i - \overline{z})^2\right)^{-1/2}$$

is then $\approx \mathrm{n}(0, 1)$ under the null hypothesis (or we could use the square of this, which would be $\approx \chi^2_1$).

The likelihood ratio statistic is

$$Q_n = 2[n\hat{\alpha} + \hat{\beta} \sum z_i - \sum y_i \exp(\hat{\alpha} + \hat{\beta} z_i) - n \log(n/\sum Y_i) - n].$$

Although they are all asymptotically equivalent, as in Example 8.1 the Wald, score and likelihood ratio tests would not give the same values for a given sample size. Also, again a slightly different test results depending on whether expected or observed information is used, and how the parameters are estimated (in the observed information).

Another hypothesis that might be of interest is whether the effect of the covariate is linear. To test this we might use the model

$$f_{Y_i}(y|\alpha, \beta) = \exp(\alpha + \beta_1 z_i + \beta_2 z_i^2) \exp[-y \exp(\alpha + \beta_1 z_i + \beta_2 z_i^2)],$$

or

$$\begin{aligned} f_{Y_i}(y|\alpha, \beta) &= \exp(\alpha + \beta_1 z_i + \beta_2 z_i^2 + \beta_3 z_i^3) \\ &\quad \exp[-y \exp(\alpha + \beta_1 z_i + \beta_2 z_i^2 + \beta_3 z_i^3)]. \end{aligned}$$

In the first model we would then test the hypothesis $\beta_2 = 0$, while in the second model we would then test the hypothesis that $\beta_2 = \beta_3 = 0$. In both cases fitting either the null or alternative model would involve iterative methods. Note that in the second model the Wald, score, and likelihood ratio tests would be asymptotically χ^2_2 under the null hypothesis. $\diamond$

Exercise 8.2 Suppose $X_1, \ldots, X_n$ are iid with the density of X_i given by

$$f(x|\beta, \gamma) = \frac{\gamma}{\beta}\left(\frac{x}{\beta}\right)^{\gamma-1} \exp[-(x/\beta)^{\gamma}].$$

(This is a Weibull density, but the parameters are slightly different than in C&B, p. 629.) Give the large sample score test for testing $H_0 : \gamma = 1$ versus $H_1 : \gamma \neq 1$. For the variance use the expected information evaluated at $\gamma = 1$ and $\beta = \hat{\beta}_0$, the MLE for β when γ is fixed at 1.

Note: If I did this correctly the adjusted information is $n(1 + r_2 - r_1^2)$, where

$$r_k = \int_0^{\infty} [\log(u)]^k u e^{-u}\, du.$$

Do not worry about evaluating these integrals, which do not have closed form antiderivatives. ◇

Exercise 8.3 Suppose $Y_i = \alpha + \beta z_i + \epsilon_i$, $i = 1, \ldots, n$, where the $\epsilon_i \overset{iid}{\sim} \mathrm{n}(0, 1)$, α and β are unknown parameters, and the z_i are fixed constants. Assume that $\sum_i z_i = 0$. Consider testing $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$.

(a). Show that the score and Wald statistics (χ_1^2 versions) are both equal to $(\sum_i Y_i z_i)^2 / \sum_i z_i^2$.

(b). Show that the likelihood ratio statistic (the large sample version Q_n) is also equal to the statistic in (a).

(c). Suppose now the hypothesis is $H_0 : \beta = \alpha = 0$. Give the Wald statistic for this hypothesis.

◇

Exercise 8.4 Suppose $X_1, \ldots, X_n$ are iid Poisson(λ) and $Y_1, \ldots, Y_m$ are iid Poisson(μ), with the X_i independent of the Y_j. Give the large sample score test for $H_0 : \lambda = \mu$ versus $H_1 : \lambda \neq \mu$. (It may help to express the parameters in a different form.) ◇