Hypothesis Testing

Zhentao Shi

October 28, 2019

Notation: $\mathbf{X}$ denotes a random variable or random vector. $\mathbf{x}$ is its realization.

1 Hypothesis Testing

- A *hypothesis* is a statement about the parameter space Θ.
- The *null hypothesis* Θ_0 is a subset of Θ of interest, ideally suggested by scientific theory.
- The *alternative hypothesis* $\Theta_1 = \Theta \backslash \Theta_0$ is the complement of Θ_0.
- *Hypothesis testing* is a decision, based on the observed evidence, to accept the null hypothesis or to reject it.
- If Θ_0 is a singleton, we call it a *simple hypothesis*; otherwise we call it a *composite hypothesis*. For example, if the parameter space $\Theta = \mathbb{R}$, then $\Theta_0 = \{0\}$ (or equivalently $\theta_0 = 0$) is a simple hypothesis, while $\Theta_0 = (-\infty, 0]$ (or equivalently $\theta_0 \leq 0$) is a composite hypothesis.
- A *test function* is a mapping
$$\phi_\theta : \mathcal{X}^n \mapsto \{0, 1\},$$
where $\mathcal{X}$ is the sample space. The null hypothesis is accepted if $\phi_\theta(\mathbf{X} = \mathbf{x}) = 0$, or rejected if $\phi_\theta(\mathbf{X} = \mathbf{x}) = 1$. Notice that the test function depends on the hypothesized parameter value θ.
- The *acceptance region* is defined as $A_\phi = \{\mathbf{x} \in \mathcal{X}^n : \phi_\theta(\mathbf{x}) = 0\}$, and the *rejection region* is $R_\phi = \{\mathbf{x} \in \mathcal{X}^n : \phi_\theta(\mathbf{x}) = 1\}$.
- The *power function* of a test ϕ_θ is
$$\beta_\phi(\theta) = P(\{\phi_\theta(\mathbf{X}) = 1\}) = E(\phi_\theta(\mathbf{X})).$$
The power function measures, at a given point θ, the probability that the test function rejects the null.
- The *power* of a test for some $\theta \in \Theta_1$ is the value of $\beta_\phi(\theta)$. The *size* of the test is $\sup_{\theta \in \Theta_0} \beta_\phi(\theta)$. Notice that the definition of power depends on a θ in the alternative hypothesis Θ_1, whereas that of size is independent of θ due to the supremum over the set of null Θ_0.
- The *level* of a test is any value $\alpha \in (0, 1)$ such that $\alpha \geq \sup_{\theta \in \Theta_0} \beta_\phi(\theta)$, which is often used when it is difficult to attain the exact supremum. A test of size α is also of level α or bigger; while a test of level α must have size smaller or equal to α.

```
| decision     |  reject $H_{1}$  | reject $H_{0}$
|--------------|------------------| ---------------
| $H_{0}$ true |     correct      | Type I error
| $H_{0}$ false| Type II error    |   correct
```

- size = P(reject H_0 when H_0 is true)
- power = P(reject H_0 when H_0 is false)
- The *probability of committing Type I error* is $\beta_\phi(\theta)$ for some $\theta \in \Theta_0$.
- The *probability of committing Type II error* is $1 - \beta_\phi(\theta)$ for $\theta \in \Theta_1$.

The philosophy on hypothesis testing has been debated for centuries. At present the prevailing framework in statistics textbooks is the frequentist perspective. A frequentist views the parameter as a fixed constant, and they keep a conservative attitude about the Type I error. Only if overwhelming evidence is demonstrated shall a researcher reject the null. Under the philosophy of protecting the null hypothesis, a desirable test should have a small level. Conventionally we take $\alpha = 0.01, 0.05$ or 0.1. There can be many tests of correct size.

Example A trivial test function, $\phi_\theta(\mathbf{X}) = 1\{0 \le U \le \alpha\}$ for all $\theta \in \Theta$, where U is a random variable from a uniform distribution on $[0,1]$, has correct size α but no power. We say a test is *unbiased* if $\beta_\phi(\theta) > \alpha$ for all $\theta \in \Theta_1$. The trivial test mentioned here is not an unbiased one. On the other extreme, the trivial test function $\phi_\theta(\mathbf{X}) = 1$ for all θ has the biggest power but incorrect size.

Usually, we design a test by proposing a test statistic $T_n : \mathcal{X}^n \times \Theta \mapsto \mathbb{R}^+$ and a critical value $c_{1-\alpha}$. Given T_n and $c_{1-\alpha}$, we write the test function as

$$\phi_\theta(\mathbf{X}) = 1\{T_n(\mathbf{X}, \theta) > c_{1-\alpha}\}.$$

To ensure such a $\phi(\mathbf{x})$ has correct size, we need to figure out the distribution of T_n under the null hypothesis (called the *null distribution*), and choose a critical value $c_{1-\alpha}$ according to the null distribution and the desirable size or level α.

The concept of *level* is useful if we do not have sufficient information to derive the exact size of a test.

Example If $(X_{1i}, X_{2i})_{i=1}^n$ are randomly drawn from some unknown joint distribution, but we know the marginal distribution is $X_{ji} \sim N(\theta_j, 1)$, for $j = 1, 2$. In order to test the joint hypothesis $\theta_1 = \theta_2 = 0$, we can construct a test function

$$\phi_{\theta_1=\theta_2=0}(\mathbf{X}_1, \mathbf{X}_2) = 1\left\{\left\{\sqrt{n}\left|\overline{X}_1\right| \ge c_{1-\alpha/4}\right\} \cup \left\{\sqrt{n}\left|\overline{X}_2\right| \ge c_{1-\alpha/4}\right\}\right\},$$

where $c_{1-\alpha/4}$ is the $(1-\alpha/4)$-th quantile of the standard normal distribution. The level of this test is

$$\begin{aligned} P\left(\phi_{\theta_1=\theta_2=0}(\mathbf{X}_1, \mathbf{X}_2)\right) &\le P\left(\sqrt{n}\left|\overline{X}_1\right| \ge c_{1-\alpha/4}\right) + P\left(\sqrt{n}\left|\overline{X}_2\right| \ge c_{1-\alpha/4}\right) \\ &= \alpha/2 + \alpha/2 = \alpha. \end{aligned}$$

where the inequality follows by the *Bonferroni inequality*

$$P(A \cup B) \le P(A) + P(B).$$

(The seemingly trivial Bonferroni inequality is useful in many proofs of probability results.) Therefore, the level of $\phi(\mathbf{X}_1, \mathbf{X}_2)$ is α, but the exact size is unknown without the knowledge of the joint distribution. (Even if we know the correlation of X_{1i} and X_{2i}, putting two marginally normal distributions together does not make a jointly normal vector in general.)

Denote the class of test functions of level α as $\Psi_\alpha = \left\{\phi : \sup_{\theta\in\Theta_0} \beta_\phi(\theta) \leq \alpha\right\}$. A *uniformly most powerful test* $\phi^* \in \Psi_\alpha$ is a test function such that, for every $\phi \in \Psi_\alpha$,

$$\beta_{\phi^*}(\theta) \geq \beta_\phi(\theta)$$

uniformly over $\theta \in \Theta_1$.

Example Suppose a random sample of size 6 is generated from

$$(X_1, \ldots, X_6) \sim \text{i.i.d.} N(\theta, 1),$$

where θ is unknown. We want to infer the population mean of the normal distribution. The null hypothesis is H_0: $\theta \leq 0$ and the alternative is H_1: $\theta > 0$. All tests in

$$\Psi = \left\{1\left\{\bar{X} \geq c/\sqrt{6}\right\} : c \geq 1.64\right\}$$

has the correct level. Since $\bar{X} = N(\theta, 1/6)$, the power function for those in Ψ is

$$\begin{aligned}
\beta_\phi(\theta) &= P\left(\bar{X} \geq \frac{c}{\sqrt{6}}\right) \\
&= P\left(\frac{\bar{X}-\theta}{1/\sqrt{6}} \geq \frac{\frac{c}{\sqrt{6}}-\theta}{1/\sqrt{6}}\right) \\
&= P\left(N \geq c - \sqrt{6}\theta\right) \\
&= 1 - \Phi\left(c - \sqrt{6}\theta\right)
\end{aligned}$$

where $N = \frac{\bar{X}-\theta}{1/\sqrt{6}}$ follows the standard normal, and Φ is the cdf of standard normal. It is clear that $\beta_\phi(\theta)$ is monotonically decreasing in c. Thus the test function

$$\phi_{\theta=0}(\mathbf{X}) = 1\left\{\bar{X} \geq 1.64/\sqrt{6}\right\}$$

is the most powerful test in Ψ, as $c = 1.64$ is the lower bound that Ψ allows.

Another commonly used indicator in hypothesis testing is p-value:

$$\sup_{\theta\in\Theta_0} P\left(T_n(\mathbf{x}, \theta) \leq T_n(\mathbf{X}, \theta)\right).$$

In the above expression, $T_n(\mathbf{x}, \theta)$ is the realized value of the test statistic T_n, while $T_n(\mathbf{X}, \theta)$ is the random variable generated by $\mathbf{X}$ under the null $\theta \in \Theta_0$. The interpretation of the p-value is tricky. p-value is the probability that we observe $T_n(\mathbf{X}, \theta)$ being greater than the realized $T_n(\mathbf{x}, \theta)$ if the null hypothesis is true. p-value is *not* the probability that the null hypothesis is true. Under the frequentist perspective, the null hypothesis is either true or false, with certainty. The randomness of a test comes only from sampling, not from the hypothesis. It measures whether the dataset is consistent with the null hypothesis, or whether the evidence from the data is compatible with the null hypothesis. p-value is closely related to the corresponding test. When p-value is smaller than the specified test size α, the test rejects the null.

So far we have been talking about hypothesis testing in finite sample. The discussion and terminologies can be carried over to the asymptotic world when $n \to \infty$. If we denote the power function as $\beta_{n,\phi}(\theta)$, in which we make its dependence on the sample size n explicit, the test is of asymptotic size α if $\limsup_{n\to\infty} \beta_{n,\phi}(\theta) \leq \alpha$ for all $\theta \in \Theta_0$. A test is *consistent* if $\beta_{n,\phi}(\theta) \to 1$ for all $\theta \in \Theta_1$.

2 Confidence Interval

An *interval estimate* is a function $C : \mathcal{X}^n \mapsto \{\Theta' : \Theta' \subseteq \Theta\}$ that maps a point in the sample space to a subset of the parameter space. The *coverage probability* of an *interval estimator* $C(\mathbf{X})$ is defined as $P_\theta(\theta \in C(\mathbf{X}))$. When θ is of one dimension, we usually call the interval estimator *confidence interval*. When θ is of multiple dimensions, we call the it *confidence region* and it of course includes the one-dimensional θ as a special case. The coverage probability is the frequency that the interval estimator captures the true parameter that generates the sample (From the frequentist perspective, the parameter is fixed while the confidence region is random). It is *not* the probability that θ is inside the given confidence interval (From the Bayesian perspective, the parameter is random while the confidence region is fixed conditional on $\mathbf{X}$.)

Exercise: Suppose a random sample of size 6 is generated from

$$(X_1, \ldots, X_6) \sim \text{i.i.d. } N(\theta, 1).$$

Find the coverage probability of the random interval is

$$\left[\bar{X} - 1.96/\sqrt{6}, \bar{X} + 1.96/\sqrt{6}\right].$$

Hypothesis testing and confidence region are closely related. Sometimes it is difficult to directly construct the confidence region, but easy to test a hypothesis. One way to construct confidence region is by *inverting a test*. Suppose ϕ_θ is a test of size α. If $C(\mathbf{X})$ is constructed as

$$C(\mathbf{X}) = \{\theta \in \Theta : \phi_\theta(\mathbf{X}) = 0\}.$$

For any $\theta \in \Theta_0$, its coverage probability

$$P(\theta \in C(\mathbf{X})) = P(\{\phi_\theta(\mathbf{X}) = 0\}) = 1 - P(\{\phi_\theta(\mathbf{X}) = 1\}) = 1 - \beta_\phi(\theta) \geq 1 - \alpha$$

where the last inequality follows as $\beta_\phi(\theta) \leq \alpha$. If Θ_0 is a singleton, the equality holds.

3 Bayesian Credible Set

The Bayesian framework offers a coherent and natural language for statistical decision. However, the major criticism against Bayesian statistics is the arbitrariness of the choice of the prior.

In the Bayesian framework, both the data $\mathbf{X}_n$ and the parameter θ are random variables. Before she observes the data, she holds a *prior distribution* π about θ. After observing the data, she updates the prior distribution to a *posterior distribution* $p(\theta|\mathbf{X}_n)$. The *Bayes Theorem* connects the prior and the posterior as

$$p(\theta|\mathbf{X}_n) \propto f(\mathbf{X}_n|\theta)\pi(\theta)$$

where $f(\mathbf{X}_n|\theta)$ is the likelihood function.

Here is a classical example to illustrate the Bayesian approach of statistical inference. Suppose we have an iid sample $\mathbf{X}_n = (X_1, \ldots, X_n)$ drawn from a normal distribution with unknown θ and known σ. If a researcher's prior distribution $\theta \sim N(\theta_0, \sigma_0^2)$, her posterior distribution is, by some routine calculation, also a normal distribution

$$p(\theta|\mathbf{x}) \sim N\left(\tilde{\theta}, \tilde{\sigma}^2\right),$$

where $\tilde{\theta} = \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\theta_0 + \frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\bar{x}$ and $\tilde{\sigma}^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}$. Thus the Bayesian credible set is

$$\left(\tilde{\theta} - z_{1-\alpha/2} \cdot \tilde{\sigma}, \tilde{\theta} + z_{1-\alpha/2} \cdot \tilde{\sigma}\right).$$

This posterior distribution depends on θ_0 and σ_0^2 from the posterior. When the sample size is sufficiently large the posterior can be approximated by $N(\bar{x}, \sigma^2/n)$, where the prior information is overwhelmed by the information accumulated from the data.

In contrast, a frequentist will estimate $\hat{\theta} = \bar{x} \sim N(\theta, \sigma^2/n)$. Her confidence interval is

$$\left(\bar{x} - z_{1-\alpha/2} \cdot \sigma/\sqrt{n}, \bar{x} - z_{1-\alpha/2} \cdot \sigma/\sqrt{n}\right).$$

For any finite n, the Bayesian credible set and the frequentist confidence interval are different.

4 Application in OLS

4.1 Wald Test

Suppose the OLS estimator $\widehat{\beta}$ is asymptotic normal, i.e.

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} N(0, \Omega)$$

where Ω is a $K \times K$ positive definite covariance matrix and R is a $q \times K$ constant matrix, then $R\sqrt{n}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} N(0, R\Omega R')$. Moreover, if rank $(R) = q$, then

$$n\left(\widehat{\beta} - \beta\right)' R' \left(R\Omega R'\right)^{-1} R\left(\widehat{\beta} - \beta\right) \xrightarrow{d} \chi_q^2.$$

Now we intend to test the null hypothesis $R\beta = r$. Under the null, the Wald statistic

$$W_n = n\left(R\widehat{\beta} - r\right)' \left(R\widehat{\Omega}R'\right)^{-1} \left(R\widehat{\beta} - r\right) \xrightarrow{d} \chi_q^2$$

where $\widehat{\Omega}$ is a consistent estimator of Ω.

Example (Single test) In a linear regression

$$y = x_i'\beta + e_i = \sum_{k=1}^{5} \beta_k x_{ik} + e_i.$$
$$E\left[e_i x_i\right] = \mathbf{0}_5,$$

where y is wage and

$$x = \left(\text{edu}, \text{age}, \text{experience}, \text{experience}^2, 1\right)'.$$

To test whether *education* affects *wage*, we specify the null hypothesis $\beta_1 = 0$. Let $R = (1, 0, 0, 0, 0)$ and $r = 0$.

$$\sqrt{n}\widehat{\beta}_1 = \sqrt{n}\left(\widehat{\beta}_1 - \beta_1\right) = \sqrt{n}R\left(\widehat{\beta} - \beta\right) \xrightarrow{d} N\left(0, R\Omega R'\right) \sim N\left(0, \Omega_{11}\right), \tag{1}$$

where Ω_{11} is the $(1,1)$ (scalar) element of Ω. Under $H_0 : R\beta = (1, 0, 0, 0, 0)\begin{pmatrix}\beta_1\\ \beta_2\\ \beta_3\\ \beta_4\\ \beta_5\end{pmatrix} = \beta_1 = 0$, we have

$$\sqrt{n}R\left(\widehat{\beta} - \beta\right) = \sqrt{n}\widehat{\beta}_1 \xrightarrow{d} N\left(0, \Omega_{11}\right)$$

Therefore,

$$\sqrt{n}\frac{\widehat{\beta}_1}{\widehat{\Omega}_{11}^{1/2}} = \sqrt{\frac{\Omega_{11}}{\widehat{\Omega}_{11}}}\sqrt{n}\frac{\widehat{\beta}_1}{\sqrt{\Omega_{11}}}$$

If $\widehat{\Omega} \xrightarrow{p} \Omega$, then $\left(\Omega_{11}/\widehat{\Omega}_{11}\right)^{1/2} \xrightarrow{p} 1$ by the continuous mapping theorem. As $\sqrt{n}\widehat{\beta}_1/\Omega_{11}^{1/2} \xrightarrow{d} N\left(0, 1\right)$, we conclude $\sqrt{n}\widehat{\beta}_1/\widehat{\Omega}_{11}^{1/2} \xrightarrow{d} N\left(0, 1\right)$.

The above example is a test about a single coefficient, and the test statistic is essentially a t-statistic. The following example gives a test about a joint hypothesis.

Example (Joint test) We want to simultaneously test $\beta_1 = 1$ and $\beta_3 + \beta_4 = 2$ in the above example. The null hypothesis can be expressed in the general form $R\beta = r$, where the restriction matrix R is

$$R = \begin{pmatrix} 1 & 0 & 0 & 0 & 0\\ 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

and $r = (1, 2)'$. Once we figure out R, it is routine to construct the test.

These two examples are linear restrictions. In order to test a nonlinear regression, we need the so-called *delta method*.

Delta method If $\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} N\left(0, \Omega_{K\times K}\right)$, and $f : \mathbb{R}^K \mapsto \mathbb{R}^q$ is a continuously differentiable function for some $q \leq K$, then

$$\sqrt{n}\left(f\left(\widehat{\theta}\right) - f\left(\theta_0\right)\right) \xrightarrow{d} N\left(0, \frac{\partial f}{\partial \theta}\left(\theta_0\right)\Omega\frac{\partial f}{\partial \theta}\left(\theta_0\right)'\right).$$

This result can be easily shown by a mean-value expansion

$$f(\hat{\theta}) - f(\theta_0) = \frac{\partial f(\tilde{\theta})}{\partial \theta}(\hat{\theta} - \theta_0)$$

where $\tilde{\theta}$ lies on the line segment connecting $\hat{\theta}$ and θ_0. Multiply both sides by $\sqrt{n}$ and notice $\tilde{\theta} \xrightarrow{p} \theta_0$, by Slutsky theorem we have $\sqrt{n}(f(\hat{\theta}) - f(\theta_0)) \xrightarrow{d} \frac{\partial f}{\partial \theta}\left(\theta_0\right) N(0, \Omega)$.

In the example of linear regression, the optimal experience level can be found by setting to zero the first order condition with respective to experience, $\beta_3 + 2\beta_4\text{experience}^* = 0$. We test the hypothesis that the optimal experience level is 20 years; in other words,

$$\text{experience}^* = -\frac{\beta_3}{2\beta_4} = 20.$$

This is a nonlinear hypothesis. If $q \leq K$ where q is the number of restrictions, we have

$$n\left(f\left(\widehat{\theta}\right)-f\left(\theta_0\right)\right)'\left(\frac{\partial f}{\partial \theta}\left(\theta_0\right)\Omega\frac{\partial f}{\partial \theta}\left(\theta_0\right)'\right)^{-1}\left(f\left(\widehat{\theta}\right)-f\left(\theta_0\right)\right)\xrightarrow{d}\chi^2_q,$$

where in this example, $\theta=\beta$, $f\left(\beta\right)=-\beta_3/\left(2\beta_4\right)$. The gradient

$$\frac{\partial f}{\partial \beta}\left(\beta\right)=\left(0,0,-\frac{1}{2\beta_4},\frac{\beta_3}{2\beta_4^2},0\right)$$

Since $\widehat{\beta}\xrightarrow{p}\beta_0$, by the continuous mapping theorem, if $\beta_{0,4}\neq 0$, we have $\frac{\partial}{\partial\beta}f\left(\widehat{\beta}\right)\xrightarrow{p}\frac{\partial}{\partial\beta}f\left(\beta_0\right)$. Therefore, the (nonlinear) Wald test is

$$W_n=n\left(f\left(\widehat{\beta}\right)-20\right)'\left(\frac{\partial f}{\partial \beta}\left(\widehat{\beta}\right)\widehat{\Omega}\frac{\partial f}{\partial \beta}\left(\widehat{\beta}\right)'\right)^{-1}\left(f\left(\widehat{\beta}\right)-20\right)\xrightarrow{d}\chi^2_1.$$

This is a valid test with correct asymptotic size.

However, we can equivalently state the null hypothesis as $\beta_3+40\beta_4=0$ and we can construct a Wald statistic accordingly. Asymptotically equivalent though, in general a linear hypothesis is preferred to a nonlinear one, due to the approximation error in the delta method under the null and more importantly the invalidity of the Taylor expansion under the alternative. It also highlights the problem of Wald test being *variant* to re-parametrization.

4.2 Lagrangian Multiplier Test*

Restricted least square

$$\min_{\beta}\left(y-X\beta\right)'\left(y-X\beta\right)\text{ s.t. }R\beta=r.$$

Turn it into an unrestricted problem

$$L\left(\beta,\lambda\right)=\frac{1}{2n}\left(y-X\beta\right)'\left(y-X\beta\right)+\lambda'\left(R\beta-r\right).$$

The first-order condition

$$\begin{aligned}\frac{\partial}{\partial\beta}L&=-\frac{1}{n}X'\left(y-X\tilde{\beta}\right)+\tilde{\lambda}R=-\frac{1}{n}X'e+\frac{1}{n}X'X\left(\tilde{\beta}-\beta^*\right)+R'\tilde{\lambda}=0.\\ \frac{\partial}{\partial\lambda}L&=R\tilde{\beta}-r=R\left(\tilde{\beta}-\beta^*\right)=0\end{aligned}$$

Combine these two equations into a linear system,

$$\begin{pmatrix}\widehat{Q} & R'\\ R & 0\end{pmatrix}\begin{pmatrix}\tilde{\beta}-\beta^*\\ \tilde{\lambda}\end{pmatrix}=\begin{pmatrix}\frac{1}{n}X'e\\ 0\end{pmatrix},$$

where $\hat{Q}=X'X/n$.

Thus we can explicitly express the estimator as

$$\begin{aligned}\begin{pmatrix}\tilde{\beta}-\beta^*\\ \tilde{\lambda}\end{pmatrix}&=\begin{pmatrix}\widehat{Q} & R'\\ R & 0\end{pmatrix}^{-1}\begin{pmatrix}\frac{1}{n}X'e\\ 0\end{pmatrix}\\ &=\begin{pmatrix}\widehat{Q}^{-1}-\widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1}R\widehat{Q}^{-1} & \widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1}\\ \left(R\widehat{Q}^{-1}R'\right)^{-1}R\widehat{Q}^{-1} & \left(R'Q^{-1}R\right)^{-1}\end{pmatrix}\begin{pmatrix}\frac{1}{n}X'e\\ 0\end{pmatrix}.\end{aligned}$$

We conclude that

$$\sqrt{n}\tilde{\lambda} = \left(R\widehat{Q}^{-1}R'\right)^{-1} R\widehat{Q}^{-1}\frac{1}{\sqrt{n}}X'e \xrightarrow{d} N\left(0, \left(RQ^{-1}R'\right)^{-1} RQ^{-1}\Omega Q^{-1}R'\left(RQ^{-1}R'\right)^{-1}\right).$$

Let $W = \left(RQ^{-1}R'\right)^{-1} RQ^{-1}\Omega Q^{-1}R'\left(RQ^{-1}R'\right)^{-1}$, we have

$$n\tilde{\lambda}'W^{-1}\tilde{\lambda} \xrightarrow{d} \chi_q^2.$$

If homoskedastic, then $W = \sigma^2\left(RQ^{-1}R'\right)^{-1} RQ^{-1}QQ^{-1}R'\left(RQ^{-1}R'\right)^{-1} = \sigma^2\left(RQ^{-1}R'\right)^{-1}$. Replace W with the estimated $\hat{W}$,

$$\begin{aligned}\frac{n\tilde{\lambda}'R\hat{Q}^{-1}R'\tilde{\lambda}}{\hat{\sigma}^2} &= \frac{1}{n\hat{\sigma}^2}\left(y - X\tilde{\beta}\right)' X\hat{Q}^{-1}R'(R\hat{Q}^{-1}R')^{-1}R\hat{Q}^{-1}X'\left(y - X\tilde{\beta}\right) \\ &= \frac{1}{n\hat{\sigma}^2}\left(y - X\tilde{\beta}\right)' P_{X\hat{Q}^{-1}R'}\left(y - X\tilde{\beta}\right).\end{aligned}$$

4.3 Likelihood-Ratio test*

For likelihood ratio test, the starting point can be a criterion function $L(\beta) = (y - X\beta)'(y - X\beta)$. It does not have to be the likelihood function.

$$\begin{aligned}L\left(\tilde{\beta}\right) - L\left(\widehat{\beta}\right) &= \frac{\partial L}{\partial \beta}\left(\widehat{\beta}\right) + \frac{1}{2}\left(\tilde{\beta} - \widehat{\beta}\right)' \frac{\partial L}{\partial \beta \partial \beta}\left(\dot{\beta}\right)\left(\tilde{\beta} - \widehat{\beta}\right) \\ &= 0 + \frac{1}{2}\left(\tilde{\beta} - \widehat{\beta}\right)' \widehat{Q}\left(\tilde{\beta} - \widehat{\beta}\right).\end{aligned}$$

From the derivation of LM test, we have

$$\begin{aligned}\sqrt{n}\left(\tilde{\beta} - \beta^*\right) &= \left(\widehat{Q}^{-1} - \widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1} R\widehat{Q}^{-1}\right)\frac{1}{\sqrt{n}}X'e \\ &= \frac{1}{\sqrt{n}}\left(X'X\right)X'e - \widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1} R\widehat{Q}^{-1}\frac{1}{\sqrt{n}}X'e \\ &= \sqrt{n}\left(\widehat{\beta} - \beta^*\right) - \widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1} R\widehat{Q}^{-1}\frac{1}{\sqrt{n}}X'e\end{aligned}$$

Therefore

$$\sqrt{n}\left(\tilde{\beta} - \widehat{\beta}\right) = -\widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1} R\widehat{Q}^{-1}\frac{1}{\sqrt{n}}X'e$$

and

$$\begin{aligned}&n\left(\tilde{\beta} - \beta\right)' \widehat{Q}\left(\tilde{\beta} - \widehat{\beta}\right) \\ &= \frac{1}{\sqrt{n}}e'X\widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1} R\widehat{Q}^{-1}\widehat{Q}\widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1} R\widehat{Q}^{-1}\frac{1}{\sqrt{n}}X'e \\ &= \frac{1}{\sqrt{n}}e'X\widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1} R\widehat{Q}^{-1}\frac{1}{\sqrt{n}}X'e\end{aligned}$$

In general, it is a quadratic form of normal distributions. If homoskedastic, then

$$\left(R\widehat{Q}^{-1}R'\right)^{-1/2} R\widehat{Q}^{-1}\frac{1}{\sqrt{n}}X'e$$

has variance

$$\sigma^2 \left(RQ^{-1}R'\right)^{-1/2} RQ^{-1}QQ^{-1}R' \left(RQ^{-1}R'\right)^{-1/2} = \sigma^2 I_q.$$

We can view the optimization of the log-likelihood as a two-step optimization with the inner step $\sigma = \sigma(\beta)$. By the envelop theorem, when we take derivative with respect to β, we can ignore the indirect effect of $\partial\sigma(\beta)/\partial\beta$.