Chapter 8

Hypothesis Testing

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

A *hypothesis* is a statement about the parameter space Θ. Hypothesis testing checks whether the data support a *null hypothesis* Θ_0, which is a subset of Θ of interest. Ideally the null hypothesis should be suggested by scientific theory. The *alternative hypothesis* $\Theta_1 = \Theta \backslash \Theta_0$ is the complement of Θ_0. Based on the observed evidence, hypothesis testing decides to accept the null hypothesis or to reject it. If the null hypothesis is rejected by the data, the data is incompatible with the proposed scientific theory.

8.1 Decision Rule and Errors

8.1.1 Terminologies

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, whereas $\Theta_0 = (-\infty, 0]$ (or equivalently $\theta_0 \leq 0$) is a composite hypothesis.

A *test function* is a mapping

$$\phi_\theta : \mathcal{X}^n \mapsto \{0, 1\},$$

Table 8.1: Decisions and States

	accept H_0 (reject H_1)	reject H_0 (accept H_1)
H_0 true (H_1 false)	correct decision	Type I error
H_0 false (H_1 true)	Type II error	correct decision

where $\mathcal{X}$ is the sample space. The null hypothesis is accepted if $\phi_\theta\left(\mathbf{X}=\mathbf{x}\right)=0$, or rejected if $\phi_\theta\left(\mathbf{X}=\mathbf{x}\right)=1$. Notice that the test function depends on the hypothesized parameter value θ. We call the set $A_\phi=\left\{\mathbf{x}\in\mathcal{X}^n:\phi_\theta\left(\mathbf{x}\right)=0\right\}$ the *acceptance region*, and its complement $R_\phi=\left\{\mathbf{x}\in\mathcal{X}^n:\phi_\theta\left(\mathbf{x}\right)=1\right\}$ the *rejection region.*

The *power function* of a test ϕ_θ is

$$\beta_\phi\left(\theta\right)=P\left(\left\{\phi_\theta\left(\mathbf{X}\right)=1\right\}\right)=E\left(\phi_\theta\left(\mathbf{X}\right)\right).$$

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 $\beta_\phi\left(\theta\right)$. The *size* of the test is $\sup_{\theta\in\Theta_0}\beta_\phi\left(\theta\right)$. 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 $\alpha\geq\sup_{\theta\in\Theta_0}\beta_\phi\left(\theta\right)$, 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.

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

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. 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, $\phi_\theta(\mathbf{X}) = 1\{0 \leq U \leq \alpha\}$ 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 enjoys the biggest power but suffers 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.

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

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

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. p-value measures whether the dataset 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 every $\theta \in \Theta_1$.

Example 8.1. The concept of *level* is useful if we do not have sufficient information to derive the exact size of a test. 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

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

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

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

where the inequality follows by the *Bonferroni inequality*

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

(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.)

8.1.2 Optimality

Just like there may be multiple valid estimators for a task of estimation, there may be multiple tests for a task of hypothesis testing. For a class of tests of the same level α under the null $\Psi_\alpha = \left\{\phi : \sup_{\theta\in\Theta_0} \beta_\phi\left(\theta\right) \leq \alpha\right\}$, it is natural to prefer a test ϕ^* that exhibits higher power than all other tests under consideration at each point of the alternative hypothesis in that

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

for every $\theta \in \Theta_1$ and every $\phi \in \Psi_\alpha$. If such a test $\phi^* \in \Psi_\alpha$ exists, we call it a *uniformly most powerful test.*

Example 8.2. Suppose a random sample of size 6 is generated from

$$\left(X_1, \ldots, X_6\right) \sim \text{iid.}N\left(\theta, 1\right),$$

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 $\bar{X} = N\left(\theta, 1/6\right)$, the power function for those in Ψ is

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

where $N = \frac{\bar{X}-\theta}{1/\sqrt{6}}$ follows the standard normal, and Φ is the cdf of the 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_{α} allows in order to keep the level α.

8.1.3 Likelihood-Ratio Test and Wilks' theorem

The mean of the Gaussian model can be written in a closed-form. When closed-forms are unavailable, the likelihood-ratio test serves as a very general testing statistic under the likelihood principle. Suppose $\widehat{\theta} = \arg\max_{\theta\in\Theta} \ell_n(\theta)$, where $\ell_n(\theta) = n^{-1}\sum_i \log f(x_i;\theta)$ is the maximum likelihood estimator (MLE). Take a Taylor expansion of $\ell_n(\theta_0)$ around $\ell_n\left(\widehat{\theta}\right)$:

$$\begin{aligned}\ell_n(\theta_0) - \ell_n\left(\widehat{\theta}\right) &= \frac{\partial \ell_n}{\partial\theta}\left(\widehat{\theta}\right)'\left(\theta_0 - \widehat{\theta}\right) + \frac{1}{2}\left(\theta_0 - \widehat{\theta}\right)'\left(\frac{\partial^2}{\partial\theta\partial\theta'}\ell_n(\dot{\theta})\right)\left(\theta_0 - \widehat{\theta}\right)\\ &= \frac{1}{2}\left(\widehat{\theta} - \theta_0\right)'\left(\frac{\partial^2}{\partial\theta\partial\theta'}\ell_n(\dot{\theta})\right)\left(\widehat{\theta} - \theta_0\right).\end{aligned}$$

where $\dot{\theta}$ lies in the line segment connecting θ_0 and $\widehat{\theta}$, and the last equality follows as $\frac{\partial \ell_n}{\partial\theta}\left(\widehat{\theta}\right) = 0$ due to the first order condition of optimality.

Define $L_n(\theta) := \sum_i \log f(x_i;\theta)$, and the *likelihood-ratio statistic* as

$$\mathcal{LR} := 2\left(L_n\left(\widehat{\theta}\right) - L_n(\theta_0)\right) = 2n\left(\ell_n\left(\widehat{\theta}\right) - \ell_n(\theta_0)\right).$$

Obviously $\mathcal{LR} \geq 0$ because $\widehat{\theta}$ maximizes $\ell_n(\theta)$. Multiply $-2n$ to the two sides of the above Taylor expansion:

$$\mathcal{LR} = \sqrt{n}\left(\widehat{\theta}-\theta_0\right)'\left(-\frac{\partial^2}{\partial\theta\partial\theta'}\ell_n\left(\dot{\theta}\right)\right)\sqrt{n}\left(\widehat{\theta}-\theta_0\right).$$

Notice that when the model is correctly specified, as $\widehat{\theta}\xrightarrow{p}\theta_0$ we have

$$-\frac{\partial^2}{\partial\theta\partial\theta'}\ell_n\left(\dot{\theta}\right)\xrightarrow{p}-\mathcal{H}\left(\theta_0\right)=\mathcal{I}\left(\theta_0\right)$$
$$\sqrt{n}\left(\widehat{\theta}-\theta_0\right)\xrightarrow{d}N\left(0,\mathcal{I}^{-1}\left(\theta_0\right)\right)$$

By Slutsky's theorem:

$$\left(-\frac{\partial^2}{\partial\theta\partial\theta'}\ell_n\left(\dot{\theta}\right)\right)^{1/2}\left[\sqrt{n}\left(\widehat{\theta}-\theta_0\right)\right]\xrightarrow{d}\mathcal{I}^{1/2}\left(\theta_0\right)\times N\left(0,\mathcal{I}^{-1}\left(\theta_0\right)\right)\sim N\left(0,I_k\right).$$

and then $\mathcal{LR}\xrightarrow{d}\chi^2\left(K\right)$ by the continuous mapping theorem. The fact that when the parametric model is correctly specified, $\mathcal{LR}\xrightarrow{d}\chi^2\left(K\right)$ is called Wilks' theorem, or Wilks' phenomenon.

8.2 Confidence Interval

An *interval estimate* is a function $C:\mathcal{X}^n\mapsto\{\Theta_1:\Theta_1\subseteq\Theta\}$ that maps a point in the sample space to a subset of the parameter space. The *coverage probability* of an *interval estimator* $C\left(\mathbf{X}\right)$ is defined as $P_\theta\left(\theta\in C\left(\mathbf{X}\right)\right)$. 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 8.1. Suppose a random sample of size 6 is generated from

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

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.

```r
set.seed(2020-10-28)
# function for two-sided confidence interval
CI <- function(x) {# x is a vector of random variables
  # nominal coverage probability is 90%
  n <- length(x)
  mu <- mean(x)
  sig <- sd(x)
  upper <- mu + 1.645 / sqrt(n) * sig
  lower <- mu - 1.645/ sqrt(n) * sig
  return(list(lower = lower, upper = upper))
}
```

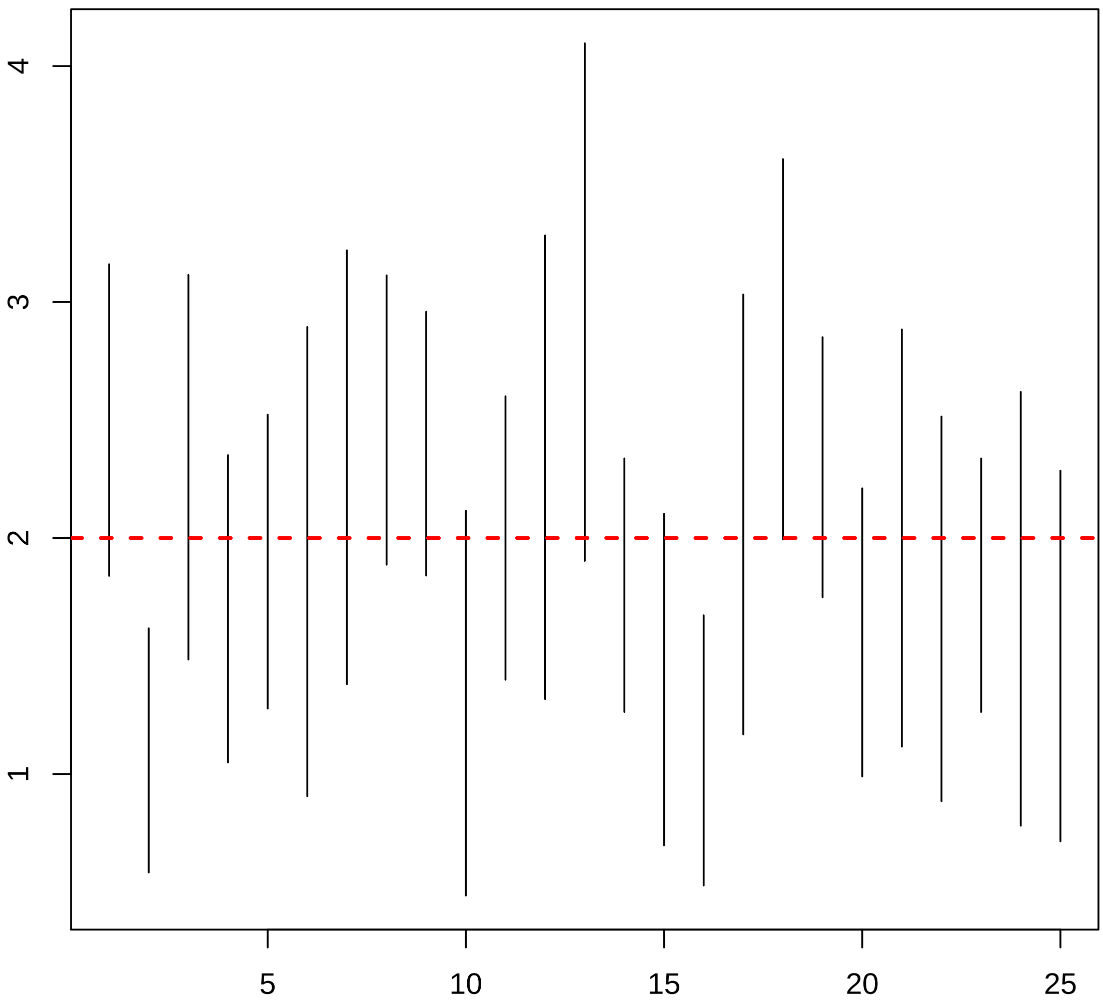
Empirical coverage probability

```
Rep <- 1000
sample_size <- 10
capture <- rep(0, Rep)
Bounds <- matrix(0, nrow = Rep, ncol = 2)
for (i in 1:Rep) {
  mu <- 2
  x <- rpois(sample_size, mu)
  bounds <- CI(x)
  capture[i] <- ((bounds$lower <= mu) & (mu <= bounds$upper))
  Bounds[i,] <- unlist( bounds )
}
cat("the emprical coverage probability = ", mean(capture)) # empirical size

## the emprical coverage probability =  0.845
```

Plot confidence intervals for 25 replications.

```
Bounds25 <- Bounds[1:25, ]
plot(1, type="n", xlab="", ylab="",
     ylim=c(min(Bounds25), max(Bounds25)), xlim=c(1, 25))
segments(x0= 1:25, y0=Bounds25[,1], x1 = 1:25, y1 = Bounds25[,2])
abline(h=2, col = "red", lty = 2, lwd = 2)
```



8.3 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 prior. 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

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

The Bayesian credible set and the frequentist confidence interval are different for any finite n, but they coincide when $n \to \infty$.

8.4 Application in OLS

8.4.1 Wald Test

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

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

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

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

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

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

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

Example 8.3. (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 $r=0$.

$$\sqrt{n}\widehat{\beta}_1=\sqrt{n}\left(\widehat{\beta}_1-\beta_1\right)=\sqrt{n}R\left(\widehat{\beta}-\beta\right)\xrightarrow{d}N\left(0,R\Omega R'\right)\sim N\left(0,\Omega_{11}\right), \tag{8.1}$$

where Ω_{11} is the $(1,1)$ (scalar) element of Ω. Under

$$H_0: R\beta=(1,0,0,0,0)\left(\beta_1,\ldots,\beta_5\right)'=\beta_1=0,$$

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 $\sqrt{n}\widehat{\beta}_1/\Omega_{11}^{1/2}\xrightarrow{d}N(0,1)$, we conclude $\sqrt{n}\widehat{\beta}_1/\widehat{\Omega}_{11}^{1/2}\xrightarrow{d}N(0,1)$.

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 8.4. (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 use the delta method.

Example 8.5. 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, β_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.

8.4.2 Lagrangian Multiplier Test

The Lagrangian multiplier test the regression coefficient is based on the restricted estimator

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

We know that the restricted minimization problem can be converted into an unrestricted problem

$$L\left(\beta, \lambda\right) = \frac{1}{2n} \left(y - X\beta\right)' \left(y - X\beta\right) + \lambda' \left(R\beta - r\right), \tag{8.2}$$

where $L\left(\beta, \lambda\right)$ is called the Lagrangian, and λ is the Lagrangian multiplier.

Set the first-order condition of (8.2) as zero:

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

where $\tilde{\beta}$ and $\tilde{\lambda}$ denote the roots of these equation, and β^* is the hypothesized true value. The two equations can be written as 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 $\hat{Q} = X'X/n$.

Fact 8.1.

$$\begin{pmatrix} \widehat{Q}^{-1} - \widehat{Q}^{-1} R' \left(R\widehat{Q}^{-1}R'\right)^{-1} R\widehat{Q}^{-1} & \widehat{Q}^{-1} R' \left(R\widehat{Q}^{-1}R'\right)^{-1} \\ \left(R\widehat{Q}^{-1}R'\right)^{-1} R\widehat{Q}^{-1} & \left(R'Q^{-1}R\right)^{-1} \end{pmatrix} \begin{pmatrix} \widehat{Q} & R' \\ R & 0 \end{pmatrix} = I_{K+q}.$$

Given the above fact, we can explicitly express

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

The $\tilde{\lambda}$ component is

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

as $\widehat{Q} \xrightarrow{p} Q$. Denote $W = \left(RQ^{-1}R'\right)^{-1}RQ^{-1}\Omega Q^{-1}R'\left(RQ^{-1}R'\right)^{-1}$, we have

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

Let $\widehat{W} = \left(R\widehat{Q}^{-1}R'\right)^{-1}R\widehat{Q}^{-1}\widehat{\Omega}\widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1}$. If $\widehat{\Omega} \xrightarrow{p} \Omega$, we have

$$\begin{aligned} \mathcal{LM} &= n\tilde{\lambda}'\widehat{W}^{-1}\tilde{\lambda} = n\tilde{\lambda}'W^{-1}\tilde{\lambda} + n\tilde{\lambda}'\left(\widehat{W}^{-1} - W^{-1}\right)\tilde{\lambda} \\ &= n\tilde{\lambda}'W^{-1}\tilde{\lambda} + o_p(1) \xrightarrow{d} \chi^2_q. \end{aligned}$$

This is the general expression of the LM test.

In the special case of homoskedasticity, $W = \sigma^2\left(RQ^{-1}R'\right)^{-1}RQ^{-1}QQ^{-1}R'\left(RQ^{-1}R'\right)^{-1} = \sigma^2\left(RQ^{-1}R'\right)^{-1}$. Replace W with the estimated $\hat{W}$, we have

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

Under heteroskedasticity, the tests must be based on efficient estimators. For regression, the GLS is the efficient estimator. It is as if we restore the homoskedasticity.

8.4.3 Likelihood-Ratio Test for Regression

In the previous section we have discussed the LRT. Here we put it into the context regression with Gaussian error. Let $\gamma = \sigma_e^2$. Under the classical assumptions of normal regression model,

$$L_n(\beta,\gamma) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\gamma - \frac{1}{2\gamma}(Y - X\beta)'(Y - X\beta).$$

For the unrestricted estimator, we know

$$\widehat{\gamma} = \gamma\left(\widehat{\beta}\right) = n^{-1}\left(Y - X\widehat{\beta}\right)'\left(Y - X\widehat{\beta}\right)$$

and

$$\widehat{L}_n = L_n\left(\widehat{\beta},\widehat{\gamma}\right) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\widehat{\gamma} - \frac{n}{2}$$

and the restricted estimator $\tilde{L}_n = L_n\left(\tilde{\beta},\tilde{\gamma}\right) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\tilde{\gamma} - \frac{n}{2}$. The likelihood ratio is

$$\mathcal{LR} = 2\left(\widehat{L}_n - \tilde{L}_n\right) = n\log\left(\tilde{\gamma}/\widehat{\gamma}\right).$$

If the normal regression is correctly specified, we can immediately conclude $\mathcal{LR} \xrightarrow{d} \chi^2(q)$.

Now we drop the Gaussian error assumption while keep the conditional homoskedasticity. In this case, the classical results is not applicable because $L_n(\beta,\gamma)$ is not the log-likelihood function; instead it is the *quasi log-likelihood function*. Notice

$$\begin{aligned}\mathcal{LR} &= n\log\left(1 + \frac{\tilde{\gamma} - \widehat{\gamma}}{\widehat{\gamma}}\right) = n\left(\log 1 + \frac{\tilde{\gamma} - \widehat{\gamma}}{\widehat{\gamma}} + o\left(\frac{\tilde{\gamma} - \widehat{\gamma}}{\widehat{\gamma}}\right)\right)\\ &= n\frac{\tilde{\gamma} - \widehat{\gamma}}{\widehat{\gamma}} + O\left(n\frac{\|\tilde{\gamma} - \widehat{\gamma}\|^2}{\widehat{\gamma}^2}\right) \qquad (8.3)\end{aligned}$$

by a Taylor expansion of $\log\left(1+\frac{\tilde{\gamma}-\widehat{\gamma}}{\widehat{\gamma}}\right)$ around $\log 1 = 0$. We focus on

$$\begin{aligned} n\left(\tilde{\gamma}-\widehat{\gamma}\right) &= n\left(\gamma\left(\tilde{\beta}\right)-\gamma\left(\widehat{\beta}\right)\right) \\ &= n\left(\frac{\partial\gamma\left(\widehat{\beta}\right)}{\partial\beta}+\frac{1}{2}\left(\tilde{\beta}-\widehat{\beta}\right)'\frac{\partial^2\gamma\left(\widehat{\beta}\right)}{\partial\beta\partial\beta'}\left(\tilde{\beta}-\widehat{\beta}\right)+o\left(\left\|\tilde{\beta}-\widehat{\beta}\right\|_2^2\right)\right) \\ &= \sqrt{n}\left(\tilde{\beta}-\widehat{\beta}\right)'\widehat{Q}\sqrt{n}\left(\tilde{\beta}-\widehat{\beta}\right)+O\left(n\left\|\tilde{\beta}-\widehat{\beta}\right\|_2^3\right) \end{aligned} \tag{8.4}$$

where the last line follows by $\frac{\partial\gamma\left(\widehat{\beta}\right)}{\partial\beta}=0$ and $\frac{1}{2}\frac{\partial^2\gamma\left(\widehat{\beta}\right)}{\partial\beta\partial\beta'}=\widehat{Q}$.

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

$$\begin{aligned} & \sqrt{n}\left(\tilde{\beta}-\widehat{\beta}\right)'\widehat{Q}\sqrt{n}\left(\tilde{\beta}-\widehat{\beta}\right) \\ = & \frac{1}{\sqrt{n}}e'X\widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1}R\widehat{Q}^{-1}\widehat{Q}\widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1}R\widehat{Q}^{-1}\frac{1}{\sqrt{n}}X'e \\ = & \frac{1}{\sqrt{n}}e'X\widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1}R\widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1}R\widehat{Q}^{-1}\frac{1}{\sqrt{n}}X'e \\ = & \frac{1}{\sqrt{n}}e'X\widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1}R\widehat{Q}^{-1}\frac{1}{\sqrt{n}}X'e. \end{aligned} \tag{8.5}$$

Collecting (8.3), (8.4) and (8.5), we have

$$\begin{aligned}\mathcal{LR} &= n\frac{\sigma_e^2}{\widehat{\gamma}}\cdot\frac{\tilde{\gamma}-\widehat{\gamma}}{\sigma_e^2}+o_p\left(1\right)\\ &= \frac{1}{\sqrt{n}}\frac{e}{\sigma_e}'X\widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1}R\widehat{Q}^{-1}\frac{1}{\sqrt{n}}X'\frac{e}{\sigma_e}+o_p\left(1\right)\end{aligned}$$

Notice that under homoskedasticity, CLT gives

$$\begin{aligned}R\widehat{Q}^{-1}\frac{1}{\sqrt{n}}X'e &= R\widehat{Q}^{-1/2}\widehat{Q}^{-1/2}\frac{1}{\sqrt{n}}X'\frac{e}{\sigma_e}\\ &\xrightarrow{d} RQ^{-1/2}\times N\left(0,I_K\right)\sim N\left(0,RQ^{-1}R'\right),\end{aligned}$$

and therefore the quadratic form $\frac{1}{\sqrt{n}}\frac{e}{\sigma_e}'X\widehat{Q}^{-1}R'\left(R\widehat{Q}^{-1}R'\right)^{-1}R\widehat{Q}^{-1}\frac{1}{\sqrt{n}}X'\frac{e}{\sigma_e}\xrightarrow{d}\chi^2\left(q\right)$. Moreover, $\frac{\sigma_e^2}{\widehat{\gamma}}\xrightarrow{p}1$. By Slutsky's theorem, we conclude

$$\mathcal{LR}\xrightarrow{d}\chi^2\left(q\right)$$

under homoskedasticity.

Under heteroskedasticity, again LRT must be based on efficient estimators.

8.5 Appendix

8.5.1 Neyman-Pearson Lemma

We have discussed an example of the uniformly most power test in the Gaussian location model. Under the likelihood principle, if the test is a simple versus simple (the null hypothesis is a singleton θ_0 and the alternative hypothesis is another single point θ_1), then likelihood ratio test (LRT)

$$\phi\left(\mathbf{X}\right):=1\left\{\mathcal{LR}\geq c_{LR}\right\},$$

where c_{LR} is the critical value depending on the size of the the test, is a uniformly most powerful test. This result is the well known Neyman-Pearson Lemma.

Notice $\exp(L_n(\theta)) = \Pi_i f(x_i;\theta) = f(\mathbf{x};\theta)$ where $f(\mathbf{x};\theta_0)$ is the joint density of $(x_1,\ldots,x_n)$, the LRT can be equivalently written in likelihood ratio form (without log)

$$\phi(\mathbf{X}) = 1\left\{f\left(\mathbf{X};\widehat{\theta}\right)/f(\mathbf{X};\theta_0) \geq c\right\}$$

where $c := \exp(c_{LR}/2)$.

To see its is the most power test in the simple to simple context, consider another test ψ of the same size at the single null hypothesis $\int \phi(\mathbf{x}) f(\theta_0) = \int \psi(\mathbf{x}) f(\mathbf{x};\theta_0) = \alpha$, where $f(\mathbf{x};\theta_0) =$ is the joint density of the sample $\mathbf{X}$. For any constant $c > 0$, the power of ϕ at the alternative θ_1 is

$$\begin{aligned}
E_{\theta_1}[\phi(\mathbf{X})] &= \int \phi(\mathbf{x}) f(\mathbf{x};\theta_1) \\
&= \int \phi(\mathbf{x}) f(\mathbf{x};\theta_1) - c\left[\int \phi(\mathbf{x}) f(\mathbf{x};\theta_0) - \int \psi(\mathbf{x}) f(\mathbf{x};\theta_0)\right] \\
&= \int \phi(\mathbf{x}) f(\mathbf{x};\theta_1) - c\int \phi(\mathbf{x}) f(\mathbf{x};\theta_0) + c\int \psi(\mathbf{x}) f(\mathbf{x};\theta_0) \\
&= \int \phi(\mathbf{x})\left(f(\mathbf{x};\theta_1) - cf(\mathbf{x};\theta_0)\right) + c\int \psi(\mathbf{x}) f(\mathbf{x};\theta_0). \qquad (8.6)
\end{aligned}$$

Define $\xi_c := f(\mathbf{x};\theta_1) - cf(\mathbf{x};\theta_0)$. The fact that $\phi(\mathbf{x}) = 1$ if $\xi_c \geq 0$ and $\phi(\mathbf{x}) = 0$ if $\xi_c < 0$ implies

$$\begin{aligned}
& \int \phi(\mathbf{x})\left(f(\mathbf{x};\theta_1) - cf(\mathbf{x};\theta_0)\right) = \int \phi(\mathbf{x})\,\xi_c \\
= \;& \int_{\{\xi_c \geq 0\}} \phi(\mathbf{x})\,\xi_c + \int_{\{\xi_c < 0\}} \phi(\mathbf{x})\,\xi_c = \int_{\{\xi_c \geq 0\}} \xi_c = \int \xi_c \cdot 1\{\xi_c \geq 0\} \\
\geq \;& \int \psi(\mathbf{x})\,\xi_c \cdot 1\{\xi_c \geq 0\} = \int_{\{\xi_c \geq 0\}} \psi(\mathbf{x})\,\xi_c \\
\geq \;& \int_{\{\xi_c \geq 0\}} \psi(\mathbf{x})\,\xi_c + \int_{\{\xi_c < 0\}} \psi(\mathbf{x})\,\xi_c = \int \psi(x)\,\xi_c \\
= \;& \int \psi(\mathbf{x})\left(f(\mathbf{x};\theta_1) - cf(\mathbf{x};\theta_0)\right)
\end{aligned}$$

where the first inequality follows because the test function $0 \leq \psi(\mathbf{x}) \leq 1$ for any realization of $\mathbf{x}$, and where the second inequality holds because $\int_{\{\xi_c<0\}} \psi(\mathbf{x})\, \xi_c \leq 0$. We continue (8.6):

$$\begin{aligned} E_{\theta_1}[\phi(\mathbf{X})] &\geq \int \psi(\mathbf{x})\left(f(\mathbf{x};\theta_1) - cf(\mathbf{x};\theta_0)\right) + c\int \psi(\mathbf{x}) f(\mathbf{x};\theta_0) \\ &= \int \psi(\mathbf{x}) f(\mathbf{x};\theta_1) = E_{\theta_1}[\psi(\mathbf{X})]. \end{aligned}$$

In other words, $\phi(\mathbf{X})$ is more powerful at θ_1 than any other test ψ of the same size at the null.

Neyman-Pearson lemma establishes the optimality of LRT in single versus simple hypothesis testing. It can be generalized to show the existence of the uniformly most power test in one sided composite null hypothesis $H_0 : \theta \leq \theta_0$ or $H_0 : \theta \geq \theta_0$ in the parametric class of distributions exhibiting *monotone likelihood ratio*.

`Zhentao Shi. November 2, 2020`