12. Hypothesis Testing

- introduction
- Wald test
- likelihood-based tests
- significance test for linear regression

Introduction

elements of statistical tests

- null hypothesis, alternative hypothesis
- test statistics
- rejection region
- type of errors: type I and type II errors
- confidence intervals, p-values

examples of hypothesis tests:

- hypothesis tests for the mean, and for comparing the means
- hypothesis tests for the variance, and for comparing variances

Testing procedures

a test consists of

- providing a statement of the hypotheses (H_0 (null) and H_1 (alternative))
- giving a rule that dictates if H_0 should be rejected or not

the decision rule involves a test statistic calculated on observed data

the Neyman-Pearson methodology partitions the sample space into two regions

the set of values of the test statistic for which:

the null hypothesis is rejected	**rejection region**
we fail to reject the null hypothesis	**acceptance region**

Test errors

since a test statistic is random, the same test can lead to different conclusions

- **type I error**: the test leads to *reject* H_0 when it is *true*
- **type II error:** the test *fails* to reject H_0 when it is *false*; sometimes called false alarm

probabilities of the errors:

- let β be the probability of type II error
- the **size** of a test is the probability of a type I error and denoted by α
- the **power** of a test is the probability of rejecting a false H_0 or $(1 - \beta)$

α is known as **significance level** and typically controlled by an analyst

for a given α, we would like β to be as small as possible

Some common tests

- normal test
- t-test
- F-test
- Chi-square test

e.g. a test is called a t-test if the test statistic follows t-distribution

two approaches of hypothesis test

- critical value approach
- p-value approach

Critical value approach

Definition: the critical value (associated with a significance level α) is the value of the known distribution of the test statistic such that the probability of type I error is α

steps involved this test

1. define the null and alternative hypotheses.

2. assume the null hypothesis is true and calculate the value of the test statistic

3. set a small significance level (typically $\alpha = 0.01, 0.05$, or 0.10) and determine the corresponding critical value

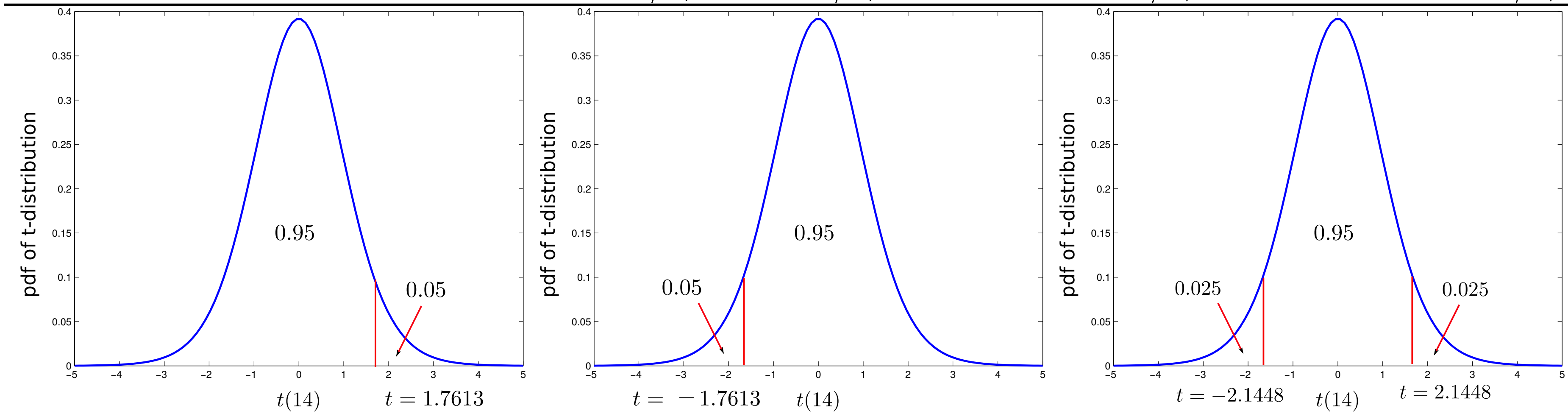
4. compare the test statistic to the critical value

condition	decision
the test statistic is more extreme than the critical value	reject H_0
the test statistic is less extreme than the critical value	accept H_0

example: hypothesis test on the population mean

- samples $N = 15$, $\alpha = 0.05$
- the test statistic is $t^* = \frac{\bar{x} - \mu}{s/\sqrt{N}}$ and has t-distribution with $N - 1$ df

test	H_0	H_1	critical value	reject H_0 if
right-tail	$\mu = 3$	$\mu > 3$	$t_{\alpha,N-1}$	$t^* \geq t_{\alpha,N-1}$
left-tail	$\mu = 3$	$\mu < 3$	$-t_{\alpha,N-1}$	$t^* \leq -t_{\alpha,N-1}$
two-tail	$\mu = 3$	$\mu \neq 3$	$-t_{\alpha/2,N-1}, t_{\alpha/2,N-1}$	$t^* \geq t_{\alpha/2,N-1}$ or $t^* \leq -t_{\alpha/2,N-1}$



p-value approach

Definition: the p-value is the probability that we observe a more extreme test statistic in the direction of H_1

steps involved this test

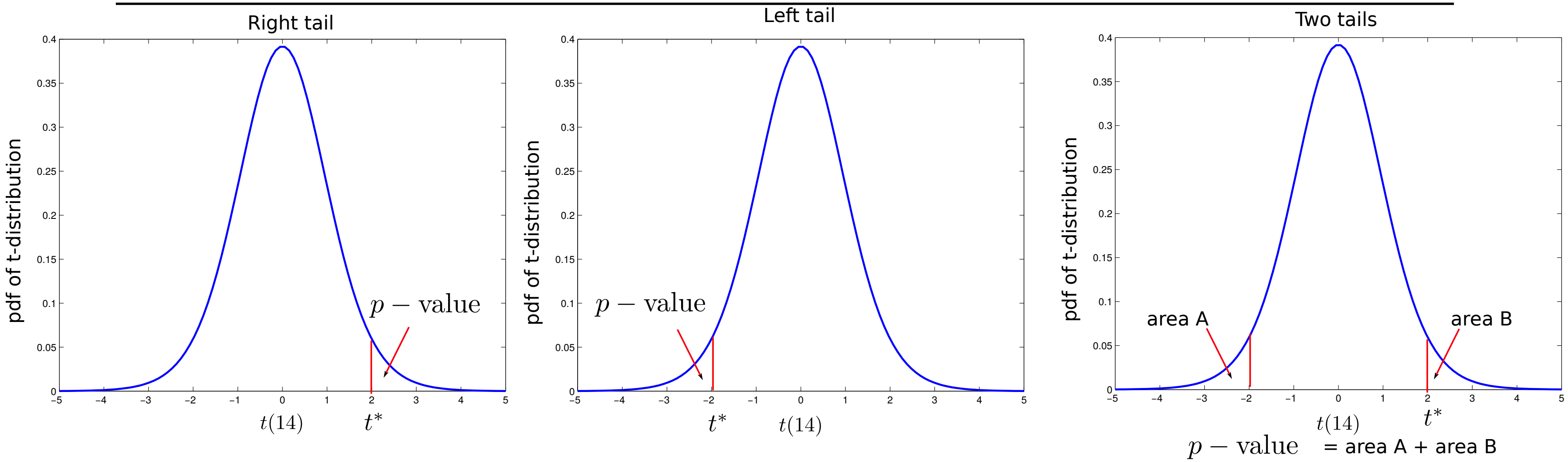
1. define the null and alternative hypotheses.
2. assume the null hypothesis is true and calculate the value of the test statistic
3. calculate the p-value using the known distribution of the test statistic
4. set a significance level α (small value such as $0.01, 0.05$)
5. compare the p-value to α

condition	decision
p-value $\leq \alpha$	reject H_0
p-value $\geq \alpha$	accept H_0

example: hypothesis test on the population mean (same as on page 12-7)

- samples $N = 15$, $\alpha = 0.01$ (have only a 1% chance of making a Type I error)
- suppose the test statistic (calculated from data) is $t^* = 2$

test	H_0	H_1	p-value expression	p-value
right-tail	$\mu = 3$	$\mu > 3$	$P(t_{14} \geq 2)$	0.0127
left-tail	$\mu = 3$	$\mu < 3$	$P(t_{14} \leq -2)$	0.0127
two-tail	$\mu = 3$	$\mu \neq 3$	$P(t_{14} \geq 2) + P(t_{14} \leq -2)$	0.0255



right-tail/left-tail tests: reject H_0, two-tail test: accept H_0

the two approaches assume H_0 were true and determine

p-value	critical value
the probability of observing a more extreme test statistic in the direction of the alternative hypothesis than the one observed	whether or not the observed test statistic is more extreme than would be expected (called critical value)

the null hypothesis is rejected if

p-value	critical value
$p - \text{value} \leq \alpha$	test statistic $\geq$ critical value

Hypothesis testing

in this chapter, we discuss about the following tests

- Wald test
- likelihood ratio test
- Lagrange multiplier (or score) test

Wald test

requires estimation of the unrestricted model

- linear hypotheses in linear models
- some Wald test statistics
- examples

Linear hypotheses in linear models

a generalization of tests for linear strictions in the linear regression model

null and alternative hypotheses for a two-sided test of linear restrictions on the regression parameters in the model: $y = X\beta + u$ are

$$\begin{aligned} H_0 &: \quad R\beta^\star - b = 0 \\ H_1 &: \quad R\beta^\star - b \neq 0 \end{aligned}$$

where $R \in \mathbf{R}^{m \times n}$ of full rank m, $\beta \in \mathbf{R}^n$ and $m \leq n$

for example, one can test $\beta_1 = 1$ and $\beta_2 - \beta_3 = 2$

the Wald test of $R\beta^\star - b = 0$ is a test of closeness to zero of the sample analogue $R\hat{\beta} - b$ where $\hat{\beta}$ is the **unrestricted OLS estimator**

assumption: suppose $u \sim \mathcal{N}(0, \sigma^2 I)$ then

$$\hat{\beta} \sim \mathcal{N}(\beta^\star, \sigma^2 (X^T X)^{-1}) \quad \Rightarrow \quad R\hat{\beta} - b \sim \mathcal{N}(0, \sigma^2 R (X^T X)^{-1} R^T)$$

under H_0 where $R\beta^\star - b = 0$

define

$$\hat{u} = y - X\hat{\beta}_{\mathrm{ls}}, \quad \mathrm{RSS} = \sum_{i=1}^{N} \hat{u}_i^2, \quad s^2 = \mathrm{RSS}/(N-n) = (N-n)^{-1} \sum_{i=1}^{N} \hat{u}_i^2$$

Facts:

- s^2 is an unbiased estimate for σ^2
- $(N-n)s^2/\sigma^2 \sim \chi^2(N-n)$

Some Wald test statistics on linear models

- known variance σ^2 (cannot be calculated in practice)

$$W_1 = (R\hat{\beta} - b)^T(\sigma^2 R(X^T X)^{-1} R^T)^{-1}(R\hat{\beta} - b) \quad \sim \quad \chi^2(m) \quad \text{under } H_0$$

- replace σ^2 by any consistent estimate s^2 (not necessarily s^2 on page 12-14)

$$W_2 = (R\hat{\beta} - b)^T(s^2 R(X^T X)^{-1} R^T)^{-1}(R\hat{\beta} - b) \quad \overset{a}{\sim} \quad \chi^2(m) \quad \text{under } H_0$$

- use $s^2 = (N-n)^{-1} \sum_i \hat{u}_i^2$

$$W_3 = (1/m)(R\hat{\beta} - b)^T(s^2 R(X^T X)^{-1} R^T)^{-1}(R\hat{\beta} - b) \;\sim\; F(m, N-n) \quad \text{under } H_0$$

simple proof:

- W_1 is in the form of $z^T A^{-1} z$ where $\mathbf{cov}(z) = A$, then

$$W_1 = (A^{-1/2} z)^T (A^{-1/2} z) \triangleq \text{ quadratic form of standard Gaussian vector}$$

  use the result on page 3-51: quadratic form of Gaussian is Chi-square

- $W_2 = (\sigma^2/s^2) W_1$ and $\mathbf{plim}(\sigma^2/s^2) = 1$, so W_2 converges to a Chi-square *asymptotically* (use Tranformation theorem on page 5-15)

- we can write W_3 as a ratio between two scaled Chi-square RVs

$$W_3 = \frac{W_1/m}{s^2/\sigma^2} = \frac{W_1/m}{((N-n)s^2/\sigma^2)/(N-n)}$$

Wald test of one restriction

for a test of **one** restriction on linear regression model:

$$y = X\beta + u, \quad u \sim \mathcal{N}(0, \sigma^2 I) \;\; \text{homoskedasticity and } X \text{ is deterministic}$$

the hypotheses are

$$H_0 : a^T\beta - b = 0, \quad H_1 : a^T\beta - b \neq 0$$

where $a \in \mathbf{R}^n$ and $b \in \mathbf{R}$

for the LS estimate, it's easy to show that $a^T\hat{\beta} - b$ is Gaussian with

$$\mathbf{E}[a^T\hat{\beta} - b] = 0, \quad \mathbf{cov}(a^T\hat{\beta} - b) = a^T\sigma^2(X^TX)^{-1}a$$

under H_0

therefore, we can propose two Wald statistics

- **Wald z-test statistic:**

$$W_4 = \frac{a^T\hat{\beta} - b}{\sqrt{a^T\sigma^2(X^TX)^{-1}a}} \sim \mathcal{N}(0,1)$$

- **Wald t-test statistic:** use $s^2 = \mathrm{RSS}/(N-n)$

$$W_5 = \frac{a^T\hat{\beta} - b}{\sqrt{a^T s^2(X^TX)^{-1}a}} \sim t_{N-n}$$

we can write W_5 a ratio of standard normal to sqrt of scaled Chi-square:

$$W_5 = \frac{a^T\hat{\beta} - b}{\sqrt{a^T s^2(X^TX)^{-1}a}} = \frac{\frac{a^T\hat{\beta}-b}{\sqrt{a^T\sigma^2(X^TX)^{-1}a}}}{\sqrt{\frac{\frac{(N-n)s^2}{\sigma^2}}{(N-n)}}} \sim t_{N-n}$$

Example on one exclusion restriction

consider the exclusion restriction that β_1 is zero: $a = (1, 0, \ldots, 0), b = 0$

suppose we use $s^2 = \mathrm{RSS}/(N-n)$ so that

$$\widehat{\mathbf{Avar}}(\hat{\beta}) = s^2(X^TX)^{-1}, \quad \widehat{\mathbf{Avar}}(\hat{\beta}_1) = \big(s^2(X^TX)^{-1}\big)_{11}$$

Wald test statistics for **exclusion restriction** are

$$\begin{aligned}
W_3 &= \frac{\hat{\beta}_1^2}{\widehat{\mathbf{Avar}}(\hat{\beta}_1)} \sim F(1, N-n) \\
W_5 &= \frac{\hat{\beta}_1}{\sqrt{\widehat{\mathbf{Avar}}(\hat{\beta}_1)}} \sim t_{N-n} \\
W_2 &= \frac{\hat{\beta}_1^2}{\widehat{\mathbf{Avar}}(\hat{\beta}_1)} \overset{a}{\sim} \chi^2(1) \quad \text{if another consistent } s^2 \text{ is used}
\end{aligned}$$

Nonlinear hypotheses

consider hypothesis tests of m restriction that are **nonlinear** in θ

let $\theta \in \mathbf{R}^n$ and $r(\theta) : \mathbf{R}^n \to \mathbf{R}^m$ be **restriction function**

the null and alternative hypotheses for a two-sided tests are

$$H_0 : r(\theta^\star) = 0, \quad H_1 : r(\theta^\star) \neq 0$$

examples: $r(\theta) = \theta_2 = 0$ or $r(\theta) = \frac{\theta_1}{\theta_2} - 1 = 0$

assumptions:

- the Jacobian matrix of r: $R(\theta) = Dr(\theta) \in \mathbf{R}^{m \times n}$ is full rank m at $\theta^\star$
- parameters are not at the boundary of Θ under H_0, *e.g.*, we rule out

$$H_0 : \theta_1 = 0 \quad \text{if the model requires } \theta_1 \geq 0$$

Wald test statistic for nonlinear restriction

intuition: obtain $\hat{\theta}$ w/o imposing restrictions and see if $r(\hat{\theta}) \approx 0$

the **Wald test statistic**

$$W = r(\hat{\theta})^T [R(\hat{\theta}) \widehat{\mathbf{Avar}}(\hat{\theta}) R(\hat{\theta})^T]^{-1} r(\hat{\theta})$$

is *asymptotically* $\chi^2(m)$ distributed under H_0

two equivalent conditions in testing:

- H_0 is rejected against H_1 at significance level α if $W > \chi^2_\alpha(m)$
- H_0 is rejected at level α if the **p-value**: $P(\chi^2(m) > W) < \alpha$

Example of nonlinear restriction

let $\theta \in \mathbf{R}^n$ and consider a test of single nonlinear restriction

$$H_0 : r(\theta) = \theta_1/\theta_2 - 1 = 0$$

then $R(\theta) \in \mathbf{R}^{1\times n}$ and given by

$$R(\theta) = \begin{bmatrix} 1/\theta_2 & -\theta_1/\theta_2^2 & 0 & \cdots & 0 \end{bmatrix}$$

let a_{ij} be (i,j) entry of $\widehat{\mathbf{Avar}}(\hat{\theta})$

$$\begin{aligned} W &= \left(\frac{\theta_1}{\theta_2} - 1\right)^2 \left(\begin{bmatrix} \frac{1}{\theta_2} & -\frac{\theta_1}{\theta_2^2} & \mathbf{0} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots \\ a_{21} & a_{22} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} \frac{1}{\theta_2} \\ -\frac{\theta_1}{\theta_2^2} \\ \mathbf{0} \end{bmatrix} \right)^{-1} \\ &= [\theta_2(\theta_1 - \theta_2)]^2 \left(\theta_2^2 a_{11} - 2\theta_1\theta_2 a_{12} + \theta_1^2 a_{22}\right)^{-1} \overset{a}{\sim} \chi^2(1) \end{aligned}$$

(θ is evaluated at $\hat{\theta}$ and the sample size N is hidden in a_{ij})

Derivation of the Wald statistic

assumption: $\hat{\theta}$ has a normal limit distribution:

$$\sqrt{N}(\hat{\theta} - \theta^\star) \xrightarrow{d} \mathcal{N}(0, P)$$

proof: starting from the first-order Taylor expansion of r under H_0

- expand r around $\theta^\star$

$$r(\hat{\theta}) = r(\theta^\star) + \nabla r(\zeta)(\hat{\theta} - \theta^\star), \quad \zeta \text{ is between } \hat{\theta} \text{ and } \theta^\star$$

- write $\sqrt{N}(r(\hat{\theta}) - r(\theta^\star)) = R(\zeta)\sqrt{N}(\hat{\theta} - \theta^\star)$ and note that

$$R(\zeta) \xrightarrow{p} R(\theta^\star), \quad \sqrt{N}(\hat{\theta} - \theta^\star) \xrightarrow{d} \mathcal{N}(0, P)$$

(use that R is continuous, apply Slustky and sandwich theorems)

- by Product Limit Normal Rule on page 5-17

$$\sqrt{N}(r(\hat{\theta}) - r(\theta^\star)) \xrightarrow{d} \mathcal{N}(0, R(\theta^\star)PR(\theta^\star)^T)$$

- under H_0: $r(\theta^\star) = 0$ and use $\mathbf{Avar}(\hat{\theta}) = P/N$, we can write

$$r(\hat{\theta}) \xrightarrow{d} \mathcal{N}(0, R(\theta^\star)\,\mathbf{Avar}(\hat{\theta})R(\theta^\star)^T)$$

- a quadratic form of standard Gaussian is a chi-square (on page 3-51)

$$r(\hat{\theta})^T \left[R(\theta^\star)\,\mathbf{Avar}(\hat{\theta})R(\theta^\star)^T\right]^{-1} r(\hat{\theta}) \quad \overset{a}{\sim} \quad \chi^2(m)$$

- the Wald test statistic is obtained by using estimates of $R(\theta^\star)$ and $\mathbf{Avar}(\hat{\theta})$

$$R(\hat{\theta}), \quad \widehat{\mathbf{Avar}}(\hat{\theta}) = \hat{P}/N$$

Likelihood-based tests

hypothesis testings when the likelihood function is known

- Wald test
- likelihood ratio (LR) test
- Lagrange multiplier (or score) test

we denote

- $L(\theta) = f(y_1, \ldots, y_N | x_1, \ldots, x_N, \theta)$ – likelihood function
- $r(\theta) : \mathbf{R}^n \rightarrow \mathbf{R}^m$ restriction function with $H_0 : r(\theta) = 0$
- $\hat{\theta}_u$: unrestricted MLE which maximizes L
- $\hat{\theta}_r$: restricted MLE which maximizes the Lagrangian $\log L(\theta) - \lambda^T r(\theta)$

Likelihood ratio test

idea: if H_0 is true, the unconstrained and constrained maximization of $\log L$ should be the same

it can be shown that the **likelihood ratio test statistic**:

$$\mathrm{LR} = -2[\log L(\hat{\theta}_r) - \log L(\hat{\theta}_u)]$$

is *asymptotically* chi-square distributed under H_0 with degree of freedom m

- if H_0 is true, $r(\hat{\theta}_u)$ should be close to zero
- note that $\log L(\hat{\theta}_u)$ is always greater than $\log L(\hat{\theta}_r)$
- LR test requires both $\hat{\theta}_u$ and $\hat{\theta}_r$
- m is the number of restriction equations

Wald test

idea: if H_0 is true, $\hat{\theta}_u$ should satisfy $r(\hat{\theta}_u) \approx 0$

- specifically for MLE, the estimate covariance satisfies CR bound and IM equality:

$$\mathbf{Avar}(\hat{\theta}_u) = \mathcal{I}_N(\theta)^{-1} = -\left(\mathbf{E}[\nabla^2 \log L(\theta^\star)]\right)^{-1} \quad \triangleq \quad P/N$$

- this leads to the **Wald test** statistic

$$W = r(\hat{\theta}_u)^T \left[R(\hat{\theta}_u)\widehat{\mathbf{Avar}}(\hat{\theta}_u)R(\hat{\theta}_u)^T\right]^{-1} r(\hat{\theta}_u) \quad \overset{a}{\sim} \quad \chi^2(m)$$

  where $\widehat{\mathbf{Avar}}(\hat{\theta}_u)$ is an estimated asymptotic covariance of $\hat{\theta}_u$

- the advantage over LR test is that only $\hat{\theta}_u$ is required

Lagrange multiplier (or score) test

ideas:

- we know that $\nabla \log L(\hat{\theta}_u) = 0$ (because it's an unconstrained maximization)
- if H_0 is true, then maximum should also occur at $\hat{\theta}_r$: $\nabla \log L(\hat{\theta}_r) \approx 0$
- LM test is called **score** test because $\nabla \log L(\theta)$ is the score vector

maximizing the Lagrangian: $\log L(\theta) - \lambda^T r(\theta)$ implies that

$$\nabla \log L(\hat{\theta}_r) = \nabla r(\hat{\theta}_r)^T \lambda$$

tests based on λ are equivalent to tests based on $\nabla \log L(\hat{\theta}_r)$ because we assume $\nabla r(\theta)$ to be full rank the LM test requires the asymptotic distribution of $\log L(\hat{\theta}_r)$

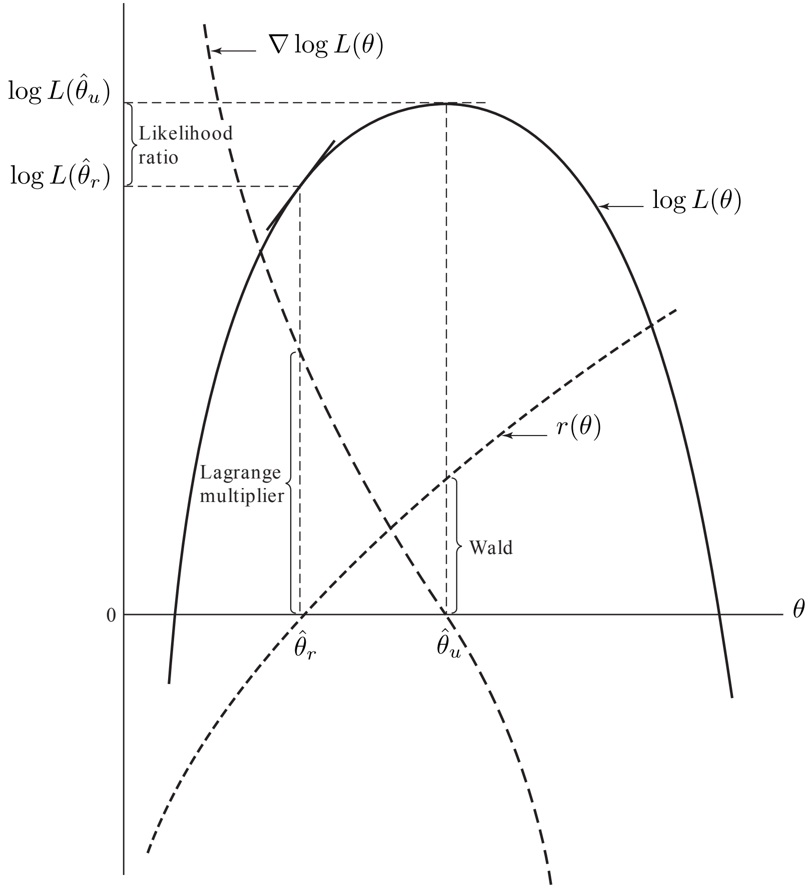
note that the asymptotic covariance of $\nabla \log L(\hat{\theta}_r)$ is the information matrix

this leads to the **Lagrange multiplier test** or **score test** statistic

$$\mathrm{LM} = (\nabla \log L(\hat{\theta}_r))^T [\mathcal{I}_N(\hat{\theta}_r)]^{-1} (\nabla \log L(\hat{\theta}_r))$$

which is asymptotically chi-square with m degree of freedoms

Graphical interpretation of loglikelihood based tests



W.H. Greene, Econometric Analysis, Prentice Hall, 2008

- Wald test checks if $r(\hat{\theta}_u) \approx 0$
- LR test checks the difference between $\log L(\hat{\theta}_u)$ and $\log L(\hat{\theta}_r)$
- LM test checks that the slope of log-likelihood at the restricted estimator should be near zero

Gaussian example

consider i.i.d. example with $y_i \in \mathcal{N}(\mu^\star, 1)$ with hypothesis

$$H_0 : \mu^\star = \mu_0, \quad (\mu_0 \text{ is just a constant value, and given})$$

therefore, $\hat{\mu}_u = \bar{y}$ and $\hat{\mu}_r = \mu_0$ (restricted solution)

$$\begin{aligned} \log L(\mu) &= -(N/2)\log 2\pi - (1/2)\sum_i (y_i - \mu)^2 \\ \nabla_\mu \log L(\mu) &= \sum_i (y_i - \mu) \end{aligned}$$

- LR test: with some algebra, we can write that

$$\mathrm{LR} = 2[\log L(\bar{y}) - \log L(\mu_0)] = N(\bar{y} - \mu_0)^2$$

- Wald test: we check if $\bar{y} - \mu_0 \approx 0$

- under H_0, the true mean of y_i is μ_0, so $\mathbf{E}[\bar{y}] = \mu_0$ and $\mathbf{var}(\bar{y}) = 1/N$
- hence, $(\bar{y} - \mu_0) \sim \mathcal{N}(0, 1/N)$

$$W = (\bar{y} - \mu_0)(1/N)^{-1}(\bar{y} - \mu_0) = N(\bar{y} - \mu_0)^2$$

- LM test: check if $\nabla \log L(\mu_0) = N(\bar{y} - \mu_0) \approx 0$

$$\nabla \log L(\mu_0) = \sum_i (y_i - \mu_0) = N(\bar{y} - \mu_0)$$

LM is just a rescaling of $(\bar{y} - \mu_0)$, so $\mathrm{LM} = W$

in conclusion, the three statistics are equivalent asymptotically

$$\mathrm{LM} = W = \mathrm{LR}$$

but they differ in finite samples

MATLAB example

the example is based on estimating the mean of a Gaussian model

```
N = 50; mu = 2; y = mu + randn(N,1); ybar = mean(y); mu0 = mu;
dof = 1; % number of restrictions

% Wald
rw = ybar-mu0; Rw = 1; EstCov = 1/N;
[h,pValue,stat,cValue] = waldtest(rw,Rw,EstCov)

% LR
uLL = -(1/2)*sum((y-ybar).^2);rLL = -(1/2)*sum((y-mu0).^2);
[h,pValue] = lratiotest(uLL,rLL,dof)

% LM
score = N*(ybar-mu0);EstCov = 1/N; % I_N(\theta) = N
[h,pValue] = lmtest(score,EstCov,dof)
```

the three tests are the same and the result is

```
ybar =

    1.9770

h =
     0

pValue =
    0.8711

stat =
    0.0263

cValue =
    3.8415
```

the p-value is greater than $\alpha = 0.05$ (default value), so H_0 is accepted

Poisson regression example

consider the log-likelihood function in the Poisson regression model

$$\log L(\beta) = \sum_{i=1}^{N} [-e^{x_i^T \beta} + y_i x_i^T \beta - \log y_i!]$$

suppose $\beta = (\beta_1, \beta_2)$ and $H_0 : r(\beta) = \beta_2 = 0$

the first and second derivatives of $\log L(\beta)$ are

$$\nabla \log L(\beta) = \sum_i (y_i - e^{x_i^T \beta}) x_i, \quad \nabla^2 \log L(\beta) = -\sum_i e^{x_i^T \beta} x_i x_i^T$$

- unrestricted MLE, $\hat{\beta}_u = (\hat{\beta}_{u1}, \hat{\beta}_{u2})$, satisfies $\nabla \log L(\beta) = 0$
- restricted MLE, $\hat{\beta}_r = (\hat{\beta}_{r1}, 0)$ where $\hat{\beta}_{r1}$ solves $\sum_i (y_i - e^{x_{i1}^T \beta_1}) x_{i1} = 0$

all the there statistics can be derived as

- LR test: calculate the fitted log-likelihood of $\hat{\beta}_u$ and $\hat{\beta}_r$
- Wald test:
  - compute the asympototic covariance of $\hat{\beta}_u$ and its estimate

$$\mathbf{Avar}(\hat{\beta}_u) = \mathcal{I}_N(\beta)^{-1} = -\mathbf{E}[\nabla^2 \log L(\beta)]^{-1} = \left(\mathbf{E}[\sum_i e^{x_i^T\beta} x_i x_i^T]\right)^{-1}$$

$$\widehat{\mathbf{Avar}}(\hat{\beta}_u) = \left(\sum_i e^{x_i^T\hat{\beta}_u} x_i x_i^T\right)^{-1}$$

  - from $r(\beta) = \hat{\beta}_2$ and $R(\beta) = \begin{bmatrix} 0 & I \end{bmatrix}$, Wald statistic is

$$W = \hat{\beta}_{u2}^T \left(\widehat{\mathbf{Avar}}(\hat{\beta}_u)_{22}\right)^{-1} \hat{\beta}_{u2}$$

where $\widehat{\mathbf{Avar}}(\hat{\beta}_u)_{22}$ denotes the $(2,2)$ block of $\widehat{\mathbf{Avar}}(\hat{\beta}_u)$

- LM test:
  - it is based on $\nabla \log L(\hat{\beta}_r)$

$$\nabla \log L(\hat{\beta}_r) = \sum_i (y_i - e^{x_i^T \hat{\beta}_r}) x_i = \sum_i x_i \hat{u}_i \quad \text{where} \quad \hat{u}_i = y_i - e^{x_{i1}^T \hat{\beta}_{r1}}$$

  - the LM statistic is

$$\mathrm{LM} = \left[\sum_i x_i \hat{u}_i\right]^T \left[\sum_i e^{x_{i1}^T \hat{\beta}_{r1}} x_i x_i^T\right]^{-1} \left[\sum_i x_i \hat{u}_i\right]$$

  - some further simplification is possible since $\sum_i x_{i1} \hat{u}_i = 0$ (from first-order condition)
  - LM test here is based on the correlation between the omitted regressors and the residual, $\hat{u}$

MATLAB example

data generation and solve for unrestricted and restricted MLE estimates

```
% Data generation
beta = [1 0]'; % The true value
x = randn(N,2); y = zeros(N,1);
lambda = zeros(N,1);
for k=1:N,
    lambda(k) = exp(x(k,:)*beta);
    y(k) = poissrnd(lambda(k)); % generate samples of y
end

% minimization of -Loglikelihood function (change the sign of LogL)
negLogFun = @(beta) -sum(-exp( sum(x.*repmat(beta',N,1),2) ) ...
+y.*sum(x.*repmat(beta',N,1),2)  );
beta0 = [2 2]'; % initial value
[beta_u,uLogL] = fminunc(negLogFun,beta0);
uLogL = -uLogL; % change back the sign of LogL
```

```
% solving for restricted MLE
negLogFun_r = @(beta) -sum(-exp( sum(x(:,1).*repmat(beta(1),N,1),2))...
    +y.*sum(x(:,1).*repmat(beta(1),N,1),2)  ); % when beta2 = 0
[beta_r1,rLogL] = fminunc(negLogFun_r,beta0(1))
rLogL = -rLogL; beta_r = [beta_r1 0]';
```

The two estimates are

```
beta_u =

    1.0533
    0.0847

beta_r =

    1.0611
         0
```

Wald test

```
TMP = 0;
for ii=1:N,
    TMP = TMP + exp(x(ii,:)*beta_u)*x(ii,:)'*x(ii,:);
end
rw = beta_u(2); Rw = [0 1]; EstCovw = TMP\eye(2)
[h,pValue,stat,cValue] = waldtest(rw,Rw,EstCovw)

h =
     0
pValue =
    0.4718
stat =
    0.5177
cValue =
    3.8415
```

accept H_0 since p-value is greater than $\alpha = 0.05$

LR test

```
LR = 2*(uLogL - rLogL)
[h,pValue,stat,cValue] = lratiotest(uLogL,rLogL,dof)

LR =
    0.5074

h =
     0

pValue =
     0.4763

stat =
    0.5074

cValue =
    3.8415
```

LM test

```
uhat = y - exp(sum(x.*repmat(beta_r',N,1),2)) ;
scorei =  x.*repmat(uhat,1,2) ; scorei = scorei';
score = sum(scorei,2);

% expect of outer product of gradient of LogL
EstCovlm1 = scorei*scorei';

EstCovlm2 = 0; % expectation of Hessian of LogL
for ii=1:N,
    EstCovlm2 = EstCovlm2 + exp(x(ii,:)*beta_r)*x(ii,:)'*x(ii,:);
end
% note that EstCovlm1 and EstCovlm2 should be close to each other

% choose to use EstCovlm2
EstCovlm = EstCovlm2\eye(2); % covariance matrix of parameters
[h,pValue,stat,cValue] = lmtest(score,EstCovlm,dof)
```

```
h =
     0

pValue =
    0.4727

stat =
    0.5157

cValue =
    3.8415
```

all the three tests agree what we should accept H_0 since p-value is greater than α

if we change the true value to $\beta = (1, -1)$ then

$$\hat{\beta}_u = (1.0954, -1.0916), \quad \hat{\beta}_r = (1.3942, 0)$$

we found that all the three tests reject H_0, *i.e.*, β_2 is not close to zero

Summary

the three tests are asymptotically equivalent under H_0 but they can behave rather differently in a small sample

- LR test requires calculation of both restricted and unrestricted estimators
- Wald test requires only unrestricted estimator
- LM test requires only restricted estimator

the choice among them typically depends on the ease of computation

Hypothesis Testing

- introduction
- Wald test
- likelihood-based tests
- **significance test for linear regression**

Recap of linear regression

a linear regression model is

$$y = X\beta + u$$

homoskedasticity assumption: u_i has the same variance for all i, given by σ^2

- prediction (fitted) error: $\hat{u} := \hat{y} - y = X\hat{\beta} - y$
- residual sum of squares: $\mathrm{RSS} = \|\hat{u}\|_2^2$
- a consistent estimate of σ^2: $s^2 = \mathrm{RSS}/(N-n)$
- $(N-n)s^2 \sim \chi^2(N-n)$
- square root of s^2 is called **standard error of the regression**
- $\mathbf{Avar}(\hat{\beta}) = s^2(X^TX)^{-1}$ (can replace s^2 by any consistent $\hat{\sigma}^2$)

Common tests for linear regression

- testing a hypothesis about a coefficient

$$H_0 : \beta_k = 0 \quad \text{VS} \quad H_1 : \beta_k \neq 0$$

  we can use both t and F statistics

- testing using the fit of the regression

$$H_0 : \text{reduced model} \quad \text{VS} \quad H_1 : \text{full model}$$

  if H_0 were true, the reduced model ($\beta_k = 0$) would lead to smaller prediction error than that of the full model ($\beta_k \neq 0$)

Testing a hypothesis about a coefficient

statistics for testing hypotheses:

$$H_0 : \beta_k = 0 \quad \text{VS} \quad H_1 : \beta_k \neq 0$$

- $\dfrac{\hat{\beta}_k}{\sqrt{s^2((X^TX)^{-1})_{kk}}} \sim t_{N-n}$

- $\dfrac{(\hat{\beta}_k)^2}{s^2((X^TX)^{-1})_{kk}} \sim F_{1,N-n}$

the above statistics are Wald statistics derived on page 12-17 through 12-19

- the term $\sqrt{s^2((X^TX)^{-1})_{kk}}$ is referred to **standard error of the coefficient**
- the expression of SE can be simplified or derived in many ways (please check)
- e.g. R use t-statistic (two-tail test)

Testing using the fit of the regression

hypotheses are based on the fitting quality of reduced/full models

$$H_0 : \text{reduced model} \quad \text{VS} \quad H_1 : \text{full model}$$

reduced model: $\beta_k = 0$ and full model: $\beta_k \neq 0$

the F-statistic used in this test

$$\frac{(\text{RSS}_R - \text{RSS}_F)}{\text{RSS}_F/(N-n)} \sim F(1, N-n)$$

- RSS_R and RSS_F are the residual sum squares of reduced and full models
- RSS_R cannot be smaller than RSS_F, so if H_0 were true, then the F statistic would be zero
- e.g. `fitlm` in MATLAB use this F statistic, or in ANOVA table

MATLAB example

perform t-test using $\alpha = 0.05$ and the true parameter is $\beta = (1, 0, -1, 0.5)$

realization 1: $N = 100$

```
>> [btrue b SE pvalue2side]  =
    1.0000    1.0172    0.1087    0.0000
         0    0.1675    0.0906    0.0675
   -1.0000   -1.0701    0.1046    0.0000
    0.5000    0.5328    0.1007    0.0000
```

- $\hat{\beta}$ is close to β
- it's not clear if $\hat{\beta}_2$ is zero but the test decides $\hat{\beta}_2 = 0$
- note that all coefficients have pretty much the same SE

realization 2: $N = 10$

```
>> [btrue b SE pvalue2side]  =
    1.0000    1.0077    0.2894    0.0131
         0    0.1282    0.4342    0.7778
   -1.0000   -1.5866    0.2989    0.0018
    0.5000    0.2145    0.2402    0.4062
```

realization 3: $N = 10$

```
>> [btrue b SE pvalue2side]  =
    1.0000    0.8008    0.3743    0.0762
         0   -0.5641    0.5442    0.3399
   -1.0000   -1.1915    0.5117    0.0588
    0.5000    0.6932    0.4985    0.2137
```

- some of $\hat{\beta}$ is close to the true value but some is not
- the test 2 decides $\hat{\beta}_2$ and $\hat{\beta}_4$ are zero while the test 3 decides all β are zero
- the sample size N affects type II error (fails to reject H_0) and we get different results from different data sets

Summary

- common tests are available in many statistical softwares, e.g, `minitab`, `lm` in R, `fitlm` in MATLAB,
- one should use with care and interpret results correctly
- an estimator is random; one cannot trust its value calculated based on a data set
- examining statistical properties of an estimator is preferred

References

Chapter 12 in

J.M. Wooldridge, *Econometric Analysis of Cross Section and Panel Data*, the MIT press, 2010

Chapter 5 in

A.C. Cameron and P.K. Trivedi, *Microeconometircs: Methods and Applications*, Cambridge, 2005

Chapter 5, 11 and 16 in

W.H. Greene, *Econometric Analysis*, Prentice Hall, 2008

Review of Basic Statistics (online course)

`https://onlinecourses.science.psu.edu/statprogram`

Stat 501 (online course)

https://onlinecourses.science.psu.edu/stat501