ECON 4130
HG revised Oct 16

Lecture Notes to Rice Chapter 8

On how to handle inference in models with more than one unknown parameter when the limit distribution for the estimators is multivariate normal.

0 Introduction - summary of one-parameter (asymptotic) inference

based on estimators that are asymptotically normally distributed.

Let θ be a parameter of interest, the true value of which being unknown, in an econometric model. We wish to perform inference on θ based on a data set. Let $\hat{\theta} = \hat{\theta}_n$ be an estimator of θ based on *n* observations or *n* observation vectors. Inference on θ based on the exact distribution of $\hat{\theta}$ is only possible in some situations. More often than not we need to resort to approximate methods - e.g., asymptotic methods for large or moderately large samples or simulation techniques (e.g. bootstrap) in small or moderately large samples.

Very often (e.g. in most cases of mle, mme, ols, gls, etc estimators for cross-section data and quite often for panel- and time series data as well) we have from theory some theorem that says

$$(0\text{-}1) \quad Z_n = \sqrt{n}\frac{\hat{\theta}-\theta}{b} \quad \underset{n\to\infty}{\overset{D}{\longrightarrow}} \quad Z \sim N(0,1)$$

where $b > 0$ is some (usually unknown) constant. Basically this is all we need. For a fixed "large" *n*, we interpret (0-1) as

$$(0\text{-}2) \quad \hat{\theta} \overset{\text{approximately}}{\sim} N(\theta, \frac{b^2}{n})$$

where the expression $\frac{b^2}{n}$ is often called the "asymptotic variance" of $\hat{\theta}$ (the precise sense of which is given in (0-1)), and the square root is called the "(asymptotic) standard error", SE,

$$(0\text{-}3) \quad \text{SE}(\hat{\theta}) = \frac{b}{\sqrt{n}}$$

Now, we cannot use (0-1) or (0-2) directly for inference when *b* is unknown. The only thing we need in order to deal with this problem, however, is a consistent estimator for *b*. So let $\hat{b}$ be a consistent estimator for *b*. We then get from Slutsky's lemma (explain how) that

$$\text{(0-4)} \quad U_n = \sqrt{n}\frac{\hat{\theta}-\theta}{\hat{b}} \quad \underset{n\to\infty}{\overset{D}{\longrightarrow}} \quad Z \sim N(0,1)$$

which we, for a given "large" *n*, may interpret as

$$\text{(0-5)} \quad \hat{\theta} \overset{\text{approximately}}{\sim} N(\theta, \frac{\hat{b}_{obs}^2}{n})$$

Note that the $\hat{b}_{obs}$ in the last expression can be understood as *the observed value* of $\hat{b}$ (i.e., the estimate). In other words, for given "large" *n* we proceed *as if* the (asymptotic) standard error of $\hat{\theta}$ is known and given by

$$\text{(0-6)} \quad SE(\hat{\theta}) = \frac{\hat{b}_{obs}}{\sqrt{n}} \quad \text{(where } \hat{b}_{obs} \text{ is the estimate and } \textit{not} \text{ the estimator.)}$$

(This is the way Stata and other packages usually use the concept "standard error".)

The basis for inference can now be stated as:

For given "large" *n*

$$\text{(0-7)} \quad U_n = \sqrt{n}\frac{\hat{\theta}-\theta}{\hat{b}} = \frac{\hat{\theta}-\theta}{\text{SE}(\hat{\theta})} \overset{\text{approximately}}{\sim} N(0,1)$$

where $\hat{b}$ is a consistent estimate of *b*.

Exercise 1. Identify b and $\hat{b}$ in the following three situations:

i) $X \sim \text{binomial}(n, p), \quad \theta = p, \quad \hat{\theta} = \hat{p} = X/n$

ii) $X \sim \text{poisson}(\lambda t), \quad n = t, \quad \theta = \lambda, \quad \hat{\theta} = \hat{\lambda} = X/t$

iii) $X_1, X_2, \ldots, X_n \sim iid$ with $E(X_i) = \mu = \theta, \quad \text{var}(X_i) = \sigma^2, \quad \hat{\theta} = \hat{\mu} = \bar{X}$

From (0-7) we can now construct confidence intervals (CI) for θ, test hypotheses about θ, calculate p-values etc.:

Approximate $1-\alpha$ confidence interval for θ:

An approximate $1-\alpha$ CI for θ is obtained directly from U_n in (0-7):

Let z_p be the upper *p*-point (quantile) in *N*(0, 1), i.e., $P(Z > z_p) = p$. Then:

$$1-\alpha \approx P(-z_{\alpha/2} \le U_n \le z_{\alpha/2}) = P(-z_{\alpha/2} \le \sqrt{n}\frac{\hat{\theta}-\theta}{\hat{b}} \le z_{\alpha/2}) =$$

$$= P\left(\hat{\theta} - z_{\alpha/2}\frac{\hat{b}}{\sqrt{n}} \le \theta \le \hat{\theta} + z_{\alpha/2}\frac{\hat{b}}{\sqrt{n}}\right) =$$

$$= P\left(\hat{\theta} - z_{\alpha/2}\mathrm{SE}(\hat{\theta}) \le \theta \le \hat{\theta} + z_{\alpha/2}\mathrm{SE}(\hat{\theta})\right)$$

Hence an approximate $1-\alpha$ CI for θ is

(0-8) $\quad \hat{\theta} \pm z_{\alpha/2}\mathrm{SE}(\hat{\theta})$ or $\hat{\theta} \pm z_{\alpha/2}\frac{\hat{b}}{\sqrt{n}}$

Approximate α-level tests for θ:

Consider the three following test problems in terms of null- and alternative hypotheses:`

Problem	H_0	H_1
1	$\theta \le \theta_0$	$\theta > \theta_0$
2	$\theta \ge \theta_0$	$\theta < \theta_0$
3	$\theta = \theta_0$	$\theta \ne \theta_0$

where θ_0 is a *known* hypothetical value. We can use the same test statistic for all three problems

(0-9) $\quad W_n = \sqrt{n}\frac{\hat{\theta}-\theta_0}{\hat{b}} = \frac{\hat{\theta}-\theta_0}{\mathrm{SE}(\hat{\theta})}$

Notice the difference between U_n and W_n:

- U_n is a non observable rv (since the true value of θ is unknown) that has the same distribution ($\approx N(0,1)$) no matter if H_0 is true or false.
- W_n is an observable (its value can be calculated from the data) and its distribution is only $\approx N(0,1)$ if θ should happen to be exactly equal to θ_0.

The relationship between U_n and W_n is given by

$$W_n = \sqrt{n}\frac{\hat{\theta}-\theta_0}{\hat{b}} = \sqrt{n}\frac{\hat{\theta}-\theta+\theta-\theta_0}{\hat{b}} = U_n + \sqrt{n}\frac{\theta-\theta_0}{\hat{b}}$$

Hence, if (the true value of) $\theta > \theta_0$, W_n behaves like a $N(0, 1)$ variable plus something positive, if $\theta < \theta_0$, W_n behaves like a $N(0, 1)$ variable plus something negative, and if $\theta = \theta_0$, W_n behaves just like a $N(0, 1)$ variable.

The approximate test for problem 1 is therefore: "Reject H_0 if $W_n \geq c$", where the critical value, c, is determined by the equation $P_{\theta=\theta_0}(W_n \geq c) = \alpha$. Since $\theta = \theta_0$ implies that W_n is approximately $N(0, 1)$, we obtain $c \approx z_\alpha$ (i.e., the upper α-point in $N(0, 1)$). If w_{obs} denotes the observed value of W_n from the data, we get the (approximate) p-value: $P_{\theta=\theta_0}(W_n \geq w_{obs}) \approx P(Z \geq w_{obs}) = 1 - \Phi(w_{obs})$, where $\Phi(x)$ is the cdf of $Z \sim N(0,1)$. Similarly for the other problems summarized in table 1:

Table 1 Approximate α-level tests

Problem	H_0	H_1	Reject H_0 when	Approximate P-value
1	$\theta \leq \theta_0$	$\theta > \theta_0$	$W_n \geq z_\alpha$	$P_{\theta=\theta_0}(W_n \geq w_{obs}) \approx P(Z \geq w_{obs})$
2	$\theta \geq \theta_0$	$\theta < \theta_0$	$W_n \leq -z_\alpha$	$P_{\theta=\theta_0}(W_n \leq w_{obs}) \approx \approx P_{\theta=\theta_0}(Z \leq w_{obs})$
3	$\theta = \theta_0$	$\theta \neq \theta_0$	$\lvert W_n \rvert \geq z_{\alpha/2}$	$P_{\theta=\theta_0}(\lvert W_n \rvert \geq \lvert w_{obs} \rvert) \approx \approx P(\lvert Z \rvert \geq \lvert w_{obs} \rvert) = = 2P(Z \geq \lvert w_{obs} \rvert)$

Note that we can achieve all this based on only the estimate and standard error obtained from e.g., a computer output, and a calculator, as illustrated in the following example:

Example 0-1
Suppose that we know from some computer output the estimate (based on say 40 observations) of a certain parameter θ is $\hat{\theta}_{obs} = 1.598$ with standard error is SE=0.356. In addition we know that the theory in (0-7) applies.

Then, for example, we may calculate an approximate 95% CI for θ by

$$\hat{\theta} \pm 1.96 \cdot SE = 1.598 \pm 0.698 = [0.900,\ 2.296]$$

or an approximate 90% CI

$$\hat{\theta} \pm 1.645 \cdot SE = 1.598 \pm 0.586 = [1.012,\ 2.184]$$

Suppose we want to test $H_0 : \theta \le 1$ against $H_1 : \theta > 1$. Then $\theta_0 = 1$ above, and the test statistic is

$$W_n = \frac{\hat{\theta} - 1}{SE}$$

which gives the observed value

$$w_{obs} = \frac{1.598 - 1}{0.356} = 1.680$$

The p-value for this is approximately $P_{\theta=1}(W_n > w_{obs}) \approx 1 - \Phi(1.680) = 0.046$
Thus H_0 should be rejected on the 5% level but not on 1% level.

Most computer-packages calculate the so called "t" as well, given by $t = \frac{\hat{\theta}}{SE(\hat{\theta})}$ (equal to 4.489 in this case). Notice that *t* is the same as the test statistic W_n when $\theta_0 = 0$ in the problems above. The computer package also usually calculates the p-value for the two-sided hypothesis, $H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$, (under the somewhat cryptic heading like "p > |t|"), which corresponds to the p-value in the lower right cell of table 1. In this case the p-value in the output would be given as 0.000, which means that the p-value is less than 0.001. Thus $\hat{\theta}$ is significantly (and highly so) different from 0 in this case.

(End of example)

Inference for several parameters

In the (most common) case where θ is one of several unknown parameters in the model, exactly the same principles of inference as above for θ apply (in the case of asymptotic normality). The only difference is some technicalities in connection with the determination of standard errors of estimators from the asymptotic covariance matrix of the estimators, which will be the main topic (see section 3) in these lecture notes. As a background for this, we need a little bit of information (not in Rice) on **random vectors and matrices**, and **on the multivariate normal distribution**:

1 Random matrices

Let $Y_{ij}, \quad i = 1,2,\ldots,m, \quad j = 1,2,\ldots,n$ be random variables (r.v.'s). The matrix

$$Y = \begin{pmatrix} Y_{11} & Y_{12} & \ldots & Y_{1n} \\ Y_{21} & Y_{22} & \ldots & Y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{m1} & Y_{m2} & \ldots & Y_{mn} \end{pmatrix}$$

is called a random matrix (with a joint *mn*-dimensional distribution, $f(y_{11,}y_{12},\ldots,y_{mn})$). The expected value of *Y* is *defined* as

$$(1) \qquad \mathrm{E}(Y) \overset{\text{def}}{=} \begin{pmatrix} \mathrm{E}(Y_{11}) & \mathrm{E}(Y_{12}) & \ldots & \mathrm{E}(Y_{1n}) \\ \mathrm{E}(Y_{21}) & \mathrm{E}(Y_{22}) & \ldots & \mathrm{E}(Y_{2n}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}(Y_{m1}) & \mathrm{E}(Y_{m2}) & \ldots & \mathrm{E}(Y_{mn}) \end{pmatrix}$$

The expectation satisfies the following rules (which follows directly from the definition (1) combined with the corresponding linear properties for the expectation in the scalar case):

i. $\mathrm{E}(AY + C) = A \cdot \mathrm{E}(Y) + C$

where *A*, *C*, are any matrices of constants with dimensions compatible with *Y* (i.e. $A \sim k \times m$, and $C \sim k \times n$, where *k* is arbitrary).

ii. $\mathrm{E}(AYB + C) = A \cdot \mathrm{E}(Y) \cdot B + C$

where *A, B, C* are any constant matrices compatible with *Y* in dimension so that the product and sum is well defined..

iii. $\mathrm{E}(Y') = \left[\mathrm{E}(Y)\right]'$ where A' denotes the transposed matrix

If $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ is a n-dimensional random vector, it's expectation, $\mu = \mathrm{E}(Y)$ (sometimes written, μ_Y), is therefore the vector of individual expectations,

$$\mu = \mathrm{E}(Y) = \begin{pmatrix} \mathrm{E}(Y_1) \\ \vdots \\ \mathrm{E}(Y_n) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}$$

Let $\sigma_{ij} = \mathrm{E}\left[(Y_i - \mu_i)(Y_j - \mu_j)\right] = \sigma_{ji}$ be the covariance between Y_i and Y_j. In particular we have $\sigma_{ii} = \mathrm{E}\left[(Y_i - \mu_i)^2\right] = \mathrm{var}(Y_i)$. The covariance matrix of Y (denoted as $\mathrm{cov}(Y)$) is defined as the matrix

$$\Sigma = \begin{pmatrix} \sigma_{11} & \dots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \dots & \sigma_{nn} \end{pmatrix} = \begin{pmatrix} \mathrm{var}(Y_1) & \dots & \mathrm{cov}(Y_1, Y_n) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}(Y_n, Y_1) & \dots & \mathrm{var}(Y_n) \end{pmatrix}$$

which can be expressed as

$$\mathrm{cov}(Y) = \mathrm{E}\left[(Y - \mu)(Y - \mu)'\right] = \mathrm{E}\begin{pmatrix} Y_1 - \mu_1 \\ \vdots \\ Y_n - \mu_n \end{pmatrix}(Y_1 - \mu_1, \dots, Y_n - \mu_n) =$$

$$= \mathrm{E}\begin{pmatrix} (Y_1 - \mu_1)^2 & \dots & (Y_1 - \mu_1)(Y_n - \mu_n) \\ \vdots & \ddots & \vdots \\ (Y_n - \mu_n)(Y_1 - \mu_1) & \dots & (Y_n - \mu_n)^2 \end{pmatrix} \overset{(1)}{=} \begin{pmatrix} \sigma_{11} & \dots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \dots & \sigma_{nn} \end{pmatrix} = \Sigma$$

Example 1

Suppose that $Y_1, Y_2, \dots, Y_n$ are *iid* with expectation $\mathrm{E}(Y_i) = \eta$ and $\mathrm{var}(Y_i) = \sigma^2$. Then the vector $Y' = (Y_1, \dots, Y_n)$ has expectation

$$\mathrm{E}(Y) = \begin{pmatrix} \eta \\ \vdots \\ \eta \end{pmatrix} = \eta \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

and covariance matrix (since $\sigma_{ij} = \mathrm{cov}(Y_i, Y_j) = 0$ for $i \neq j$):

$$\mathrm{cov}(Y) = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} = \sigma^2 I_n$$

where I_n is the *n*-dimensional identity matrix. (*End of example*)

If $Y' = (Y_1, \dots, Y_n)$ is a random vector, A a $p \times n$ constant matrix, b a constant $p \times 1$ vector, we obtain from **i.-iii.** (and the fact that $(BC)' = C'B'$ for matrices B and C**):**

(2) $\mathrm{E}(AY + b) = A \cdot \mathrm{E}(Y) + b = A\mu + b$

and

(3) $\mathrm{cov}(AY + b) = A \cdot \mathrm{cov}(Y) \cdot A' = A\Sigma A'$ (i.e. a $p \times p$ matrix)

which follows from

> [remembering that for matrices *A,B,C* we always have $A(B + C) = AB + AC$ and $(B + C)A = BA + CA$ whenever the multiplication is well defined]

$$\mathrm{cov}(AY + b) = \mathrm{E}\left[(AY + b - A\mu - b)(AY + b - A\mu - b)'\right] = \mathrm{E}\left[(AY - A\mu)(AY - A\mu)'\right] =$$
$$= \mathrm{E}\left[A(Y - \mu)(Y - \mu)'A'\right] = A \cdot \mathrm{E}\left[(Y - \mu)(Y - \mu)'\right]A' = A\Sigma A'$$

In particular, if Z is a linear combination of $Y_1, \dots, Y_n$, i.e. $Z = a_0 + a_1 Y_1 + \dots + a_n Y_n$, then Z can be written, $Z = a_0 + a'Y$ where $a' = (a_1, \dots, a_n)$ and $Y = (Y_1, \dots, Y_n)'$. Then (3) gives

(4) $\mathrm{var}(Z) = \mathrm{var}(a'Y) = a'\Sigma a$ where and $\Sigma = \mathrm{cov}(Y)$.

[**Proof:** Since $Z = (a_1,\ldots,a_n)\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = a'Y$ can be considered a 1×1 matrix, we must have that $Z' = Z$, and, therefore, $\text{cov}(Z) = \text{var}(Z)$ (i.e., $\text{cov}(Z) = \text{E}\left[(Z-\text{E}(Z))(Z-\text{E}(Z))'\right] = \text{E}\left[(Z-\text{E}(Z))^2\right] = \text{var}(Z)$). We then see that (4) is a special case of (3) with $A = a'$]

Example 2 (optional reading) Ordinary least squares (OLS)

To get an idea of the power of matrix notation, consider the standard multiple regression model with one response, *Y*, and *p* explanatory variables

(5) $\quad Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + u_i \quad \text{for} \quad i = 1,2,\ldots,n$

where, for simplicity, all x_{ij} are considered fixed[1], non random quantities, and the errors, $u_1,u_2,\ldots,u_n$ are assumed to be *iid* and normally distributed with expectation, $\text{E}(u_i) = 0$ and $\text{var}(u_i) = \sigma^2$. We can write (5) in matrix form as follows

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

The three matrices on the right we denote by $X, \beta,$ and u respectively. The model can now be written as

(6) $\quad Y = X\beta + u$

where *X* is the $n\times p$ (so called) design matrix, β the $(p+1)\times 1$ vector of regression coefficients, and *u* the $n\times 1$ vector of errors. Since

[1] In the lecture note on prediction (see the appendices 1 and 2) it is shown that, for the iid models, one may often consider the explanatory variables as fixed numbers (equal to the observed values) without losing information about the unknown parameters. This makes the analysis much simpler and is often used in econometrics as a simplifying measure whenever justified.

$$\mathrm{E}(u) = \begin{pmatrix} \mathrm{E}(u_1) \\ \mathrm{E}(u_2) \\ \vdots \\ \mathrm{E}(u_n) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \underline{0}$$

(where $\underline{0}$ denotes a vector of zeroes), we get from **i.** (noting that $X\beta$ is non random)

(7) $$\mathrm{E}(Y) = X\beta + \mathrm{E}(u) = X\beta$$

The covariance matrix for *Y* becomes, since $Y - X\beta = u$, and using example 1,

(8) $$\Sigma_Y = \mathrm{cov}(Y) = \mathrm{E}\left[(Y - X\beta)(Y - X\beta)'\right] = \mathrm{E}\left[\mathrm{uu}'\right] = \mathrm{cov}(u) = \sigma^2 I_n$$

The OLS estimator, $\hat{\beta}$, for β is (by definition) obtained by minimizing the sum of squares

$$Q = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}\right)^2$$

with respect to β. Differentiating *Q* with respect to all the β_j's, and setting the derivatives equal to 0, leads to (check if you wish) the following system of equations[2] that the $\hat{\beta}_j$'s must satisfy

$$\begin{aligned}
&n\hat{\beta}_0 + \left(\textstyle\sum_i x_{i1}\right)\hat{\beta}_1 + \cdots + \left(\textstyle\sum_i x_{ip}\right)\hat{\beta}_p = \textstyle\sum_i Y_i \\
&\left(\textstyle\sum_i x_{i1}\right)\hat{\beta}_0 + \left(\textstyle\sum_i x_{i1}^2\right)\hat{\beta}_1 + \cdots + \left(\textstyle\sum_i x_{i1}x_{ip}\right)\hat{\beta}_p = \textstyle\sum_i x_{i1}Y_i \\
&\quad\cdots \\
&\left(\textstyle\sum_i x_{ip}\right)\hat{\beta}_0 + \left(\textstyle\sum_i x_{ip}x_{i1}\right)\hat{\beta}_1 + \cdots + \left(\textstyle\sum_i x_{ip}^2\right)\hat{\beta}_p = \textstyle\sum_i x_{ip}Y_i
\end{aligned}$$

Noting that the coefficients of the left side are exactly the elements in the $(p+1)\times(p+1)$ matrix $X'X$, and that the right side, written as a vector, simply is $X'Y$, we can write the system more compactly as

$$X'X\hat{\beta} = X'Y$$

[2] Often called *the normal equations* in the literature

Assuming that $X'X$ is non singular (which can be shown to be the case if no single *x*-variable can be written exactly as a linear combination of the other *x*-variables[3], which is expressed by saying that there is no *exact collinearity* between the explanatory variables), we obtain the solution (the OLS estimator)

(9) $\hat{\beta} = (X'X)^{-1}X'Y$

It is now easy to prove that $\hat{\beta}$ is unbiased since, from **i.** and (7)

(10) $\mathrm{E}(\hat{\beta}) = \mathrm{E}\left[(X'X)^{-1}X'Y\right] \overset{\text{i.}}{=} (X'X)^{-1}X'\mathrm{E}(Y) \overset{(7)}{=} (X'X)^{-1}X'X\beta = I_p\beta = \beta$

Writing $C = (X'X)^{-1}X'$, we have $\hat{\beta} = CY$, and obtain the covariance matrix from (3) and (8)

[*and also using the rule that the transposed of an inverse square matrix is the inverse of the transposed,* $\left[A^{-1}\right]' = (A')^{-1}$*, which is seen by taking the transposed of the equation,* $A \cdot A^{-1} = I$*. Remember also the* $AI_n = A$ *for any* $p \times n$*-matrix A, and that, if c is a scalar, then c as factor can be taken outside a matrix product,* $A \cdot (cB) = cAB$*.*].

$$\mathrm{cov}(\hat{\beta}) = \mathrm{cov}(CY) \overset{(3)}{=} C\Sigma_Y C' \overset{(8)}{=} C\left(\sigma^2 I_n\right)C' = \sigma^2 CC' = \sigma^2 (X'X)^{-1}X'X(X'X)^{-1}$$

Hence

(11) $\mathrm{cov}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$ (*End of example.*)

Exercise 2 (optional). (*Recommended (!) for obtaining a little bit of matrix training.*)
Suppose that we are in a situation with only one (non-random) explanatory variable, *x*, so that the model in (5) reduces to

$$Y_i = \beta_0 + \beta_1 x_i + u_i \quad \text{for} \quad i = 1, 2, \ldots, n$$

where $u_1, u_2, \ldots, u_n$ are *iid* with $E(u_i) = 0$ and $\mathrm{var}(u_i) = \sigma^2$

(a) Show that the determinant, *D*, of $X'X$ is given by

$$D = n\sum_i x_i^2 - \left(\sum_i x_i\right)^2 = n\sum_i (x_i - \bar{x})^2$$

---

[3] For example, if $x_{i5} = 3 + x_{i1} - 2x_{i2}$ should happen to be true for all $i = 1, 2, \ldots, n$ in the data, there is exact collinearity between the variables $x_1, x_2,$ and x_5. This would imply that $X'X$ is singular.

(b) Find the inverse $(X'X)^{-1}$ and conclude that

$$\text{var}(\hat{\beta}_0) = \frac{\frac{1}{n}\sum_i x_i^2}{\sum_i (x_i - \bar{x})^2}\sigma^2 \quad \text{and} \quad \text{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$$

(c) Show also by evaluating (9) that

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})Y_i}{\sum_i (x_i - \bar{x})^2} = \frac{S_{xY}}{S_x^2} \qquad \text{where}$$

$$S_{xY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y}) \quad \text{and} \quad S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

(Note that this coincides with the estimation formula used for the corresponding parameter in *no-seminar-week exercise* for week 38)

2 Multinormal distributions

We say that the vector $X' = (X_1, \ldots, X_n)$ is (multi)normally distributed with expectation $\mu = \text{E}(X)$, and covariance matrix, $\Sigma = \text{cov}(X)$ (written shortly $X \sim N(\mu, \Sigma)$), if the joint pdf is given by

(12) $$f(x_1, \ldots, x_n \mid \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)} \quad \text{where} \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \text{ and } \det(\Sigma)$$

means the determinant of Σ.

This distribution has a lot of convenient mathematical properties (see e.g. Greene, Econometric Analysis, chapter 3, for a summary), but here we only need the following:

(13)

If $X \sim N(\mu, \Sigma)$ and A is a $p \times n$ constant matrix ($p \le n$) and b a $p \times 1$ constant vector, then $Y = AX + b \sim N\big(\text{E}(Y), \text{cov}(Y)\big) \overset{(2),(3)}{=} N(A\mu + b,\ A\Sigma A')$

.

[*For proof see e.g. Greene chapter 3.*]

In particular, this shows that all marginal distributions are also normal. For example, the marginal distribution of X_1, X_2 is normal since

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = AX \text{ where } A = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \end{pmatrix} \text{ which gives (check!)}$$

$$(14) \qquad \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N(\mathrm{E}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix},\ \mathrm{cov}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}) = N(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}) \ ,$$

i.e. a bivariate normal distribution

Exercise 3. Show that the pdf in (14) as defined in (12), reduces to the bivariate normal density as defined in Example F in Rice section 3.3 (both editions). [**Hint:** Introduce the correlation, ρ, between X_1 and X_2, $\rho = \sigma_{12} \big/ \sqrt{\sigma_{11}\sigma_{22}}$, implying $\sigma_{12}^2 = \sigma_{11}\sigma_{22}\rho^2$, and the determinant, $\det(\mathrm{cov}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}) = \sigma_{11}\sigma_{22} - \sigma_{12}^2 = \sigma_{11}\sigma_{22}(1-\rho^2)$ etc.]

[**Note.** Another very convenient property of multinormal joint distributions is that the conditional distribution of any of the individual variables given values for the other variables, *is itself normally distributed* with a linear and homoscedastic regression function (general formula details are given in more advanced textbooks, e.g., Greene's book). In other words, if we can assume that all our variables are jointly normally distributed, it follows automatically that the linear and homoscedastic regression model is true (which can be estimated without loss of information by the usual OLS method). In Rice the conditional distribution for the bivariate normal case is described in example B page 148 (skipping the slightly messy algebra involved). For example, using (14) above that describes a bivariate normal distribution for $(X_1, X_2)'$, we know from above that the marginal distribution of X_1 is normal. Then, manipulating $f(x_1, x_2) / f_{X_1}(x_1)$, will after some algebra show us that the conditional distribution of X_2 given that $X_1 = x_1$, is also normally distributed ,

$(X_2 \mid x_1) \sim N\left(E(X_2 \mid x_1),\ \mathrm{var}(X_2 \mid x_1)\right)$, where $E(X_2 \mid x_1) = \mu_2 + \dfrac{\sigma_{12}}{\sigma_{11}}(x_1 - \mu_1)$ and $\mathrm{var}(X_2 \mid x_1) = \sigma_{22}(1-\rho^2)$ where $\rho = \mathrm{correlation}(X_1, X_2) = \sigma_{12} \big/ \sqrt{\sigma_{11}\sigma_{22}}$. Hence the assumption of a homoscedastic and linear regression model is automatically fulfilled. **Note** also that, if X_1, X_2 are uncorrelated (i.e., $\sigma_{12} = 0 \quad \Rightarrow \quad \rho = 0$), this implies that $(X_2 \mid x_1) \sim N\left(\mu_2, \sigma_{22}\right)$. Hence the conditional distribution does not depend on x_1 in this case, which, as we have seen in the lectures, implies that X_1 and X_2 are stochastically independent. This is another convenient property of jointly normal variables, namely that

uncorrelatedness implies independence. This is special for the normal distribution and very seldom the case for other joint distributions. In general $\rho = 0$ does not imply independence.]

Example 3 (optional) (Continuation of example 2 - optional)

The error vector, *u*, in (6) has expectation $\underline{0}$ and covariance, $\Sigma_u = \mathrm{cov}(u) = \sigma^2 I_n$. We see from (12) that saying that $u \sim N(\underline{0},\ \sigma^2 I_n)$ is the same as saying that $u_1, u_2, \ldots, u_n$ are *iid* and normally distributed with expectation, $\mathrm{E}(u_i) = 0$ and $\mathrm{var}(u_i) = \sigma^2$. In fact, we have the determinant

$$\det(\Sigma_u) = \det(\sigma^2 I_n) = \det\begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^{2n}.$$

and the exponent in (12) reduces to

$$-\frac{1}{2}(u - \mathrm{E}(u))'\Sigma_u^{-1}(u - \mathrm{E}(u)) = -\frac{1}{2}u'\left(\sigma^2 I_n\right)^{-1}u = -\frac{1}{2}u'\left(\frac{1}{\sigma^2} I_n\right)u = -\frac{1}{2\sigma^2}u'u = -\frac{1}{2\sigma^2}\sum_i u_i^2$$

Substituting in (12), shows that the joint distribution (12) reduces to the product of *n* one-dimensional $N(0, \sigma^2)$-distributions as the *iid* statement would imply. Note, in particular, that this shows that if we can assume that $u_1, u_2, \ldots, u_n$ are jointly normally distributed and uncorrelated (so that the covariance matrix is diagonal), it will imply the stronger property that they are, in fact, independent!

By (13), (7), and (8) we obtain that *Y* is normally distributed, $Y \sim N(\mathrm{E}(Y), \mathrm{cov}(Y)) = N(X\beta,\ \sigma^2 I_n)$, and, by (13) again, that $\hat{\beta} = (X'X)^{-1}X'Y$ is normally distributed

$$\hat{\beta} \sim N\left(\mathrm{E}(\hat{\beta}), \mathrm{cov}(\hat{\beta})\right) = N\left(\beta,\ \sigma^2(X'X)^{-1}\right)$$

[**Note:** For completeness sake it is worth mentioning that an unbiased (and consistent) estimator for σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\hat{u}'\hat{u} = \frac{1}{n-p-1}\sum_{i=1}^{n}\hat{u}_i^2$$

where $\hat{u} = Y - X\hat{\beta}$ is the vector of the so called *residuals*. $\hat{u}$ is an (unbiased) predictor of the non-observable error vector u. This enables us to get consistent (and unbiased in the case of non-random *X*) estimators of the covariance matrix of the regression estimators, $\hat{\beta}$.]

(*End of example*.)

3 On the asymptotic distribution for mle estimators (the multi parameter case)

In this section we will only describe how to determine the asymptotic distribution for the mle estimator in case there are several unknown parameters in the model, without going into details of derivations and proofs. A good summary of the theory (not required in this course) can be found in chapter 4 of Greene's book, *Econometric Analysis*. See also Rice at the end of section 8.5.2.

Suppose that $X_1, X_2, \ldots, X_n$ are *iid* with $X_i \sim f(x_i \mid \theta)$ (pdf), where $\theta' = (\theta_1, \theta_2, \ldots, \theta_r)$ is a *r*-dimensional vector of unknown parameters. Then the joint pdf is $\prod_{i=1}^{n} f(x_i \mid \theta)$ and the log likelihood is

$$l(\theta) = \sum_{i=1}^{n} \ln f(x_i \mid \theta)$$

The mle estimator, $\hat{\theta}$, solves r equations

$$\sum_{i=1}^{n} \frac{\partial}{\partial \theta_j} \ln f(x_i \mid \hat{\theta}) = 0, \qquad j = 1, 2, \ldots, r$$

In order to define the $r \times r$ Fisher information matrix that is needed in the asymptotic distribution of $\hat{\theta}$, we introduce

$$m_{ij}(\theta) = -\mathrm{E} \frac{\partial^2 \ln f(X_i \mid \theta)}{\partial \theta_i \partial \theta_j} \qquad i, j = 1, 2, \ldots, r$$

Then the Fisher information matrix for one observation[4] is defined as

$$I(\theta) = \begin{pmatrix} m_{11}(\theta) & \cdots & m_{1r}(\theta) \\ \vdots & \ddots & \vdots \\ m_{r1}(\theta) & \cdots & m_{rr}(\theta) \end{pmatrix}$$

Under regularity conditions similar to the one-parameter case (see, e.g., Greene for details), we have that the mle satisfies

$$\sqrt{n}(\hat{\theta} - \theta) \underset{n\to\infty}{\overset{D}{\rightarrow}} N(\underline{0}, I(\theta)^{-1})$$

The definition of convergence in distribution for random vectors is similar but slightly more technical than the definition for the one-dimensional case, and we skip the details here (see, e.g., Greene for a precise definition). However, the interpretation of the result is the same as in the one-dimensional case, i.e., that it provides us an approximate distribution for $\hat{\theta}$ for large *n*,

$$\hat{\theta} \overset{\text{approximately}}{\sim} N\left(\theta, \frac{1}{n} I(\theta)^{-1}\right)$$

Hence we can say that $\hat{\theta}$ is asymptotically unbiased with asymptotic covariance matrix, $(1/n)I(\theta)^{-1}$. This matrix is unknown since θ is unknown, but can be consistently estimated by replacing θ by $\hat{\theta}$ (or any other consistent estimator of θ)[5]. [That $\hat{\theta}$ is consistent means simply that $\hat{\theta}_j \underset{n\to\infty}{\overset{P}{\rightarrow}} \theta_j$ for all $j = 1, 2, \ldots, r$]. A generalization of Slutski's lemma to the multivariate case (details omitted), now allows us to conclude that, for large *n* (where $\hat{\theta}_{obs}$ denotes the observed value of $\hat{\theta}$)

(15) $$\boxed{\hat{\theta} \overset{\text{approximately}}{\sim} N\left(\theta, \frac{1}{n} I(\hat{\theta}_{obs})^{-1}\right)}$$

which is the important result that you should know.

Using (13) we also have the important result

---

[4] I.e, one *observation vector* if we observe several variables for each unit in the sample.

[5] The continuity property of plim is still valid in the multivariate case, and the elements in $I(\theta)^{-1}$ are continuous, which follows from the continuity of the elements of $I(\theta)$, expressing the inverse by determinants. Hence, the consistency of $\hat{\theta}$ implies the consistency of $I(\hat{\theta})^{-1}$.

(16) $$A\hat{\theta} \overset{\text{approximately}}{\sim} N\left(A\theta,\ \frac{1}{n} A \cdot I(\hat{\theta}_{obs})^{-1} A' \right)$$

for any constant, $p \times r$ matrix A.

In other words: For large fixed n we may assume that $A\hat{\theta}$ is approximately (joint) normally distributed, $A\hat{\theta} \overset{\text{approximately}}{\sim} N\left(A\theta,\ C\right)$ with *known* covariance matrix, $C = \frac{1}{n} A \cdot I(\hat{\theta}_{obs})^{-1} A'$. We then take this approximate model as our basis for inference about the unknown parameters $A\theta$ in terms of confidence intervals and hypotheses.

From this we get the following: Let $k_{ij}(\theta)$ denote element *i,j* in $I(\theta)^{-1}$. Then the estimated asymptotic variance of $\hat{\theta}_j$ is the *j*-th element on the main diagonal in the estimated covariance matrix, i.e. $k_{jj}(\hat{\theta}_{obs}) / n$.

[Follows from (16). In fact, let $a' = (0,\ldots,1,\ldots,0)$ where the 1 is in position *j* and zeroes elsewhere. Then from (16)

$$\hat{\theta}_j = a'\hat{\theta} \overset{\text{approx.}}{\sim} N\left(a'\theta, \frac{1}{n} a' I(\hat{\theta}_{obs})^{-1} a \right) = N\left(\theta_j, \frac{k_{jj}(\hat{\theta}_{obs})}{n} \right) \quad]$$

Hence, we obtain an approximate $1-\alpha$ CI for θ_j: $\hat{\theta}_j \pm z_{\alpha/2} \sqrt{k_{jj}(\hat{\theta}_{obs})} \Big/ \sqrt{n}$
where $z_{\alpha/2}$ is the upper $\alpha/2$-point in $N(0,1)$.

Example 4. Assume we want a CI for the transformed parameter, $\eta = \theta_1 - \theta_2$. This we obtain from (16): Let $b' = (1,-1,0,\ldots,0)$. Then, by (16),

$$\hat{\eta} = \hat{\theta}_1 - \hat{\theta}_2 = b'\hat{\theta} \overset{\text{approx.}}{\sim} N\left(b'\theta,\ \frac{b' I(\hat{\theta}_{obs})^{-1} b}{n} \right) = N\left(\theta_1 - \theta_2,\ \frac{1}{n}(k_{11}(\hat{\theta}_{obs}) + k_{22}(\hat{\theta}_{obs}) - 2k_{12}(\hat{\theta}_{obs})) \right)$$

which leads to the approximate $1-\alpha$ CI for $\theta_1 - \theta_2$:

$$\hat{\theta}_1 - \hat{\theta}_2 \pm z_{\alpha/2} \frac{1}{\sqrt{n}} \sqrt{k_{11}(\hat{\theta}_{obs}) + k_{22}(\hat{\theta}_{obs}) - 2k_{12}(\hat{\theta}_{obs})}$$

[Note that all covariance matrices are symmetric. Hence $k_{12}(\hat{\theta}) = k_{21}(\hat{\theta})$.]

(*End of example*.)

Example 5 (On example C in Rice section 8.5 – precipitation data)

Let X_i be the amount of precipitation for rainstorm no. *i*, $i = 1,2,\ldots,n$ ($n = 227$ observations).

Model: $X_1, X_2, \ldots, X_n$ are *iid* with $X_i \sim \Gamma(\alpha, \lambda)$. The joint distribution is

$$X_1, X_2, \ldots, X_n \sim \prod_{i=1}^{n} f(x_i \mid \alpha, \lambda) = \frac{\lambda^{n\alpha}}{\Gamma(\alpha)^n}(x_1 x_2 \cdots x_n)^{\alpha-1} e^{-\lambda \sum x_i}$$

The log likelihood is

(17) $$l(\alpha, \lambda) = n\alpha \ln\lambda + (\alpha - 1)\sum_i \ln x_i - \lambda \sum_i x_i - n \ln\Gamma(\alpha)$$

The first derivatives of *l* are

$$\frac{\partial l}{\partial \alpha} = n \ln\lambda + \sum_i \ln x_i - n \frac{\partial}{\partial \alpha} \ln\Gamma(\alpha)$$

$$\frac{\partial l}{\partial \lambda} = n \frac{\alpha}{\lambda} - \sum_i x_i$$

Setting the derivatives equal to zero and solving with respect to α and λ, gives the mle estimators $\hat{\alpha}$ and $\hat{\lambda}$.

> [**Note.** There are no explicit formulas for the solution, they must be found by numerical iterations. For example, Excel works well in this case by the Solver module: Choose two cells for the arguments α and λ, with start values e.g. at the moment estimates, and then a third cell for the function (17)[6]. Then use Solver to maximize (17). This can also be done in STATA by the ml-command, but slightly more involved.]

Using his program, Rice obtained the mle estimates.

$$\hat{\alpha} = 0,441 \quad \text{and} \quad \hat{\lambda} = 1,96$$

[6] In Excel you can calculate $\ln(\Gamma(\alpha))$ by the function GAMMALN.

We want approximate 90% CI's for α and λ based on the asymptotic normal distribution of $\hat{\alpha}$ and $\hat{\lambda}$. In order to calculate the asymptotic standard errors we need the so called di- and trigamma functions defined by:

Digamma function: $\psi(\alpha) = \frac{\partial}{\partial \alpha} \ln\Gamma(\alpha)$

Trigamma function: $\psi'(\alpha) = \frac{\partial^2}{\partial \alpha^2} \ln\Gamma(\alpha)$

Both functions can be calculated in STATA (under the names digamma and trigamma).

We need the Fisher information matrix:

$$\ln f(X_i \mid \alpha, \lambda) = \alpha \ln\lambda - \ln\Gamma(\alpha) + (\alpha - 1)\ln X_i - \lambda X_i$$

giving

$$\frac{\partial \ln f}{\partial \alpha} = \ln\lambda - \psi(\alpha) + \ln X_i \quad \text{and} \quad \frac{\partial \ln f}{\partial \lambda} = \frac{\alpha}{\lambda} - X_i$$

Hence

$$\frac{\partial^2 \ln f}{\partial \alpha^2} = -\psi'(\alpha) \quad \text{(trigamma)}$$

$$\frac{\partial^2 \ln f}{\partial \alpha \partial \lambda} = \frac{\partial^2 \ln f}{\partial \lambda \partial \alpha} = \frac{1}{\lambda}$$

$$\frac{\partial^2 \ln f}{\partial \lambda^2} = -\frac{\alpha}{\lambda^2}$$

Hence the Fisher information matrix for one observation

$$I(\alpha, \lambda) = -\mathrm{E}\begin{pmatrix} -\psi'(\alpha) & \frac{1}{\lambda} \\ \frac{1}{\lambda} & -\frac{\alpha}{\lambda^2} \end{pmatrix} = \begin{pmatrix} \psi'(\alpha) & -\frac{1}{\lambda} \\ -\frac{1}{\lambda} & \frac{\alpha}{\lambda^2} \end{pmatrix}$$

The inverse of a symmetric 2×2 matrix is

$$\begin{pmatrix} a & c \\ c & b \end{pmatrix}^{-1} = \frac{1}{ab-c^2}\begin{pmatrix} b & -c \\ -c & a \end{pmatrix}$$

Hence

$$I(\alpha,\lambda)^{-1} = \frac{1}{\frac{\alpha\psi'(\alpha)}{\lambda^2}-\frac{1}{\lambda^2}} \cdot \begin{pmatrix} \frac{\alpha}{\lambda^2} & \frac{1}{\lambda} \\ \frac{1}{\lambda} & \psi'(\alpha) \end{pmatrix} = \frac{1}{\alpha\psi'(\alpha)-1} \cdot \begin{pmatrix} \alpha & \lambda \\ \lambda & \lambda^2\psi'(\alpha) \end{pmatrix}$$

We obtain an estimate of this by substituting the mle, $\hat{\alpha} = 0,441$ and $\hat{\lambda} = 1,96$, for α and λ (skipping the index *obs* on the estimates)

$$I(\hat{\alpha},\hat{\lambda})^{-1} = \frac{1}{\hat{\alpha}\psi'(\hat{\alpha})-1}\begin{pmatrix} \hat{\alpha} & \hat{\lambda} \\ \hat{\lambda} & \hat{\lambda}^2\psi'(\hat{\alpha}) \end{pmatrix} = \begin{pmatrix} 0,25903 & 1,15123 \\ 1,15123 & 13,82770 \end{pmatrix}$$

Here we found $\psi'(\hat{\alpha}) = 6,128169$ from STATA by the command:

```
di trigamma(0.441)
```

From the theory we have that $\begin{pmatrix} \hat{\alpha} \\ \hat{\lambda} \end{pmatrix} \overset{\text{approx.}}{\sim} N\left(\begin{pmatrix} \alpha \\ \lambda \end{pmatrix},\ C\right)$, where the asymptotic covariance is

$$C = \frac{1}{n} I(\hat{\alpha},\hat{\lambda})^{-1} = \begin{pmatrix} 0,0011411 & 0,0050715 \\ 0,0050715 & 0,0609150 \end{pmatrix}$$

Hence the asymptotic standard errors

$$\text{se}(\hat{\alpha}) = \sqrt{0,0011411} = 0,03378 \text{ and } \text{se}(\hat{\lambda}) = \sqrt{0,060950} = 0,24681$$

According to the theory we then obtain approximate 90% CI for α and λ

$$\hat{\alpha} \pm 1,64 \cdot \text{se}(\hat{\alpha}) = 0,441 \pm (1,64)(0,03378) = [0,386,\ 0,496]$$

$$\hat{\lambda} \pm 1,64 \cdot \text{se}(\hat{\lambda}) = 1,96 \pm (1,64)(0,247) = [1,55,\ 2,37]$$

Rice (example E, section 8.5.3)) obtains approximate 90% CI's by the parametric bootstrap method[7]:

α: [0,404, 0,523]

λ: [1,46, 2,32]

The difference between the asymptotic intervals and the bootstrap intervals does not appear to be substantial. With as much as 227 observations it is to be expected that the asymptotic theory should work well.

[7] Not in the course curriculum for 2016.