1 September 2004

Overview of Violations of the Basic Assumptions in the Classical Normal Linear Regression Model

A. Introduction and assumptions

The classical normal linear regression model can be written as

$$y = X\beta + \varepsilon \tag{1}$$

or

$$y_t = x_t'\beta + \varepsilon_t \quad (t = 1, \ldots, n). \tag{2}$$

where x_t' is the tth row of the matrix X or simply as

$$y_t = x_t\beta + \varepsilon_t \quad (t = 1, \ldots, n). \tag{3}$$

where it is implicit that x_t is a row vector containing the regressors for the tth time period. The classical assumptions on the model can be summarized as

$$\begin{array}{ll} I & y = X\beta + \varepsilon \\ II & E(\varepsilon \mid X) = 0 \\ III & E(\varepsilon\varepsilon' \mid X) = \sigma^2 I \\ IV & X \textit{ is a nonstochastic matrix of rank } k \\ V & \varepsilon \sim N(0;\, \Sigma = \sigma^2 I) \end{array} \tag{4}$$

Assumption V as written implies II and III. These assumptions are described as

1. linearity
2. zero mean of the error vector
3. scalar covariance matrix for the error vector
4. non-stochastic X matrix of full rank
5. normality of the error vector

With normally distributed disturbances, the joint density (and therefore likelihood function) of y is

$$L(y;\mu,\Sigma=\sigma^2 I) = \frac{e^{\frac{-1}{2}(y-X\beta)'\Sigma^{-1}(y-X\beta)}}{(2\pi)^{n/2}\,|\Sigma\,|^{\frac{1}{2}}}$$

$$= \frac{e^{\frac{-1}{2\sigma^2}(y-X\beta)'(y-X\beta)}}{(2\pi)^{n/2}\,|\sigma^2 I\,|^{\frac{1}{2}}} \qquad \textbf{(5)}$$

$$= \frac{e^{\frac{-1}{2\sigma^2}(y-X\beta)'(y-X\beta)}}{(2\pi)^{n/2}\,(\sigma^2)^{\frac{n}{2}}}$$

The natural log of the likelihood function is given by

$$\ell = \log L = -\frac{(y-X\beta)'(y-X\beta)}{2\sigma^2} - \frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2$$

$$= -\frac{y'y - 2\beta' X'y + \beta' X'X\beta}{2\sigma^2} - \frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 \qquad \textbf{(6)}$$

Maximum likelihood estimators are obtained by setting the derivatives of (6) equal to zero and solving the resulting k+1 equations for the k β's and σ^2. These first order conditions for the M.L estimators are

$$\frac{\partial\ell}{\partial\beta'} = -\frac{1}{2\hat{\sigma}^2}\left(-2X'y + 2(X'X)\beta^{\ell}\right) = 0$$

$$\frac{\partial\ell}{\partial\sigma^2} = \frac{1}{2}\left(\frac{1}{\hat{\sigma}^2}\right)^2 (y-X\beta^{\ell})'(y-X\beta^{\ell}) - \frac{n}{2}\frac{1}{\sigma^2} = 0 \qquad \textbf{(7)}$$

Solving we obtain

$$\beta^{\ell} = (X'X)^{-1}X'y$$

$$\hat{\sigma}^2 = \frac{1}{n}(y-X\beta^{\ell})'(y-X\beta^{\ell}) \qquad \textbf{(8)}$$

$$= \frac{e'e}{n} = \frac{\sum_t e_t^2}{n}$$

The ordinary least squares estimator is obtained be minimizing the sum of squared errors which is defined by

$$
\begin{aligned}
SSE(\hat{\beta}) &= \sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n} (y_t - x_t'\beta)^2 \\
&= [e_1 \;\; e_2 \;\; . \;\; . \;\; . \;\; e_n] \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \\
&= e'e \\
&= (y - X\hat{\beta})'(y - X\hat{\beta}) \\
&= y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\
&= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}
\end{aligned} \tag{9}
$$

The necessary condition for $SSE(\hat{\beta})$ to be a minimum is that

$$
\begin{aligned}
\frac{\partial SSE(\hat{\beta})}{\partial \hat{\beta}'} &= 0 \\
\Rightarrow \; -2X'y + 2X'X\hat{\beta} &= 0
\end{aligned} \tag{10}
$$

This gives the normal equations which can then be solved to obtain the least squares estimator

$$
\begin{aligned}
X'X\hat{\beta} &= X'y \\
\Rightarrow \; \hat{\beta} &= (X'X)^{-1}X'y
\end{aligned} \tag{11}
$$

The maximum likelihood estimator of $\hat{\beta}$ is the same as the least squares estimator. The distribution of this estimator is given as

$$
\hat{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1}) \tag{12}
$$

We have shown that the least squares estimator is:

1. unbiased
2. minimum variance of all unbiased estimators
3. consistent
4. asymptotically normal
5. asymptotically efficient.

In this section we will discuss how the statistical properties of $\hat{\beta}$ crucially dependent upon the assumptions I-V. The discussion will proceed by dropping one assumption at a time and considering the consequences. Following a general discussion, later sections will analyze specific violations of the assumptions in detail.

B. Nonlinearity

1. nonlinearity in the variables only

If the model is nonlinear in the variables, but linear in the parameters, it can still be estimated using linear regression techniques. For example consider a set of variables $z = (z_1, z_2, \ldots, z_p)$, a set of k functions $h_1 \ldots h_k$, and parameters $\beta_1^0, \beta_2^0, \cdots, \beta_k^0$. Now define the model:

$$g(y) = \beta_1^0 h_1(z) + \beta_2^0 h_2(z) + \ldots + \beta_k^0 h_k(z) + \varepsilon \tag{13}$$

This model is linear in the parameters β^0 and can be estimated using standard techniques where the functions h_i take the place of the x variables in the standard model.

2. intrinsic linearity in the parameters

a. idea

Sometimes models are nonlinear in the parameters. Some of these may be intrinsically linear, however. In the classical model, if the k parameters $\left(\beta_1^0, \beta_2^0, \cdots, \beta_k^0\right)$ can be written as k one-to-one functions (perhaps nonlinear) of a set of k underlying parameters $\theta_1, \ldots, \theta_k$, then model is intrinsically linear in θ.

b. example

$$\begin{aligned} y_t &= A^0 x_t^{\beta_1^0} e^{\varepsilon} \\ \ln y_t &= \ln A^0 + \beta_1^0 \ln x_t + \ln \varepsilon_t \\ &= \alpha^0 + \beta_1^0 \ln x_t + \varepsilon_t \\ \alpha^0 &= \ln A^0 \end{aligned} \tag{14}$$

The model is nonlinear in the parameter A^0, but since it is linear in α^0, and α^0 is a one-to-one function of A^0, the model is intrinsically linear.

3. inherently nonlinear models

Models that are inherently nonlinear cannot be estimated using ordinary least squares and the previously derived formulas. Alternatives include Taylor's series approximations and direct nonlinear estimation. In the section on non-linear estimation we showed that the non-linear least squares estimator is:

1. consistent
2. asymptotically normal

We also showed that the maximum likelihood estimator in a general non-linear model is

1. consistent
2. asymptotically normal
3. asymptotically efficient in the sense that within the consistent asymptotic normal (CAN) class it has minimum variance

If the distribution of the error terms in the non-linear least squares model is normal, and the errors are iid(o, σ^2), then the non-linear least squares estimator and the maximum likelihood estimator will be the same, just as in the classical normal linear regression model.

C. Non-zero expected value of error term ($E(\varepsilon_t) \neq 0$)

Consider the case where ε has a non-zero expectation. The least squares estimators of β is given by

$$\hat{\beta} = (X'X)^{-1}X'y \tag{15}$$

The expected value of $\hat{\beta}$ is given as follows where $E(\varepsilon_t) = \mu_{\varepsilon_t}$

$$\begin{aligned} E(\hat{\beta}) &= (X'X)^{-1}X'E(y) \\ &= (X'X)^{-1}X'(X\beta + E(\varepsilon)) \\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\mu_{\varepsilon_t} \\ &= \beta + (X'X)^{-1}X'\mu_{\varepsilon_t} \end{aligned} \tag{16}$$

which appears to suggest that all of the least squares estimators in the vector $\hat{\beta}$ are biased.

However, if $E(\varepsilon_t) = \mu_\varepsilon$ for all t, then

$$\mu_\varepsilon = \begin{pmatrix} \mu_\varepsilon \\ \mu_\varepsilon \\ \vdots \\ \mu_\varepsilon \end{pmatrix} = \mu_\varepsilon \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \tag{17}$$

To interpret this consider the rules for matrix multiplication.

$$AB = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kn} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1k} \\ b_{21} & b_{22} & \cdots & b_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nk} \end{pmatrix}$$

$$= \left[\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kn} \end{pmatrix} \begin{pmatrix} b_{11} \\ b_{21} \\ \vdots \\ b_{n1} \end{pmatrix} \quad \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kn} \end{pmatrix} \begin{pmatrix} b_{12} \\ b_{22} \\ \vdots \\ b_{n2} \end{pmatrix} \cdots \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kn} \end{pmatrix} \begin{pmatrix} b_{1k} \\ b_{2k} \\ \vdots \\ b_{nk} \end{pmatrix} \right] \tag{18}$$

Now consider

$$(X'X)^{-1}X'X = I = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \tag{19}$$

The first column of the X matrix is a column of ones. Therefore

$$(X'X)^{-1}X' \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

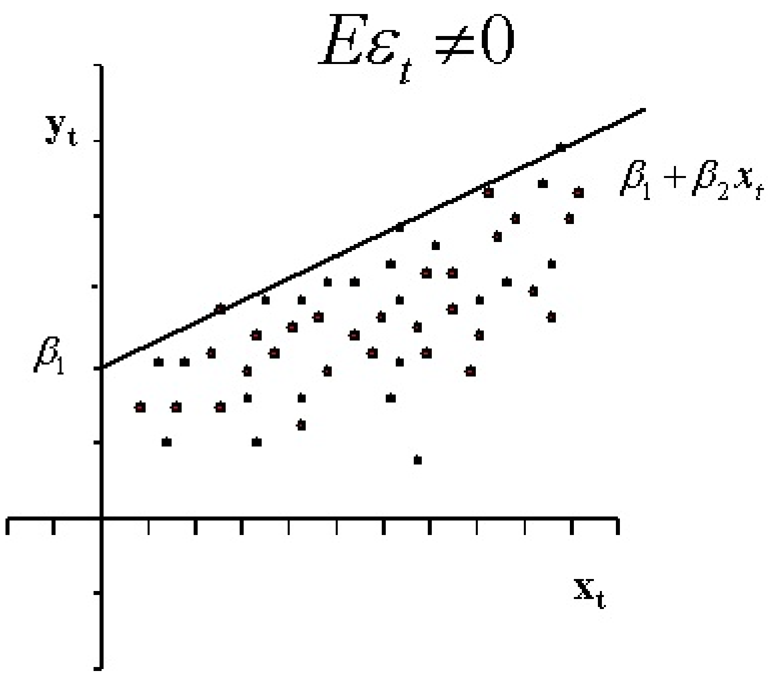
and **(20)**

$$(X'X)^{-1}X' \begin{pmatrix} \mu_\varepsilon \\ \mu_\varepsilon \\ \vdots \\ \mu_\varepsilon \end{pmatrix} = (X'X)^{-1}X' \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \mu_\varepsilon = \begin{pmatrix} \mu_\varepsilon \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Thus it is clear that

$$E(\hat{\beta}) = \begin{pmatrix} \beta_1 + \mu_\varepsilon \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_k \end{pmatrix} \tag{21}$$

and only the estimator of the intercept is biased. This situation can arise if a relevant and important factor has been omitted from the model, but the factor doesn't change over time. The effect of this variable is then included in the intercept and separate estimators of β_1 and μ_ε can't be obtained.



More general violations lead to more serious problems and in general the least squares estimators of β and σ^2 are biased.

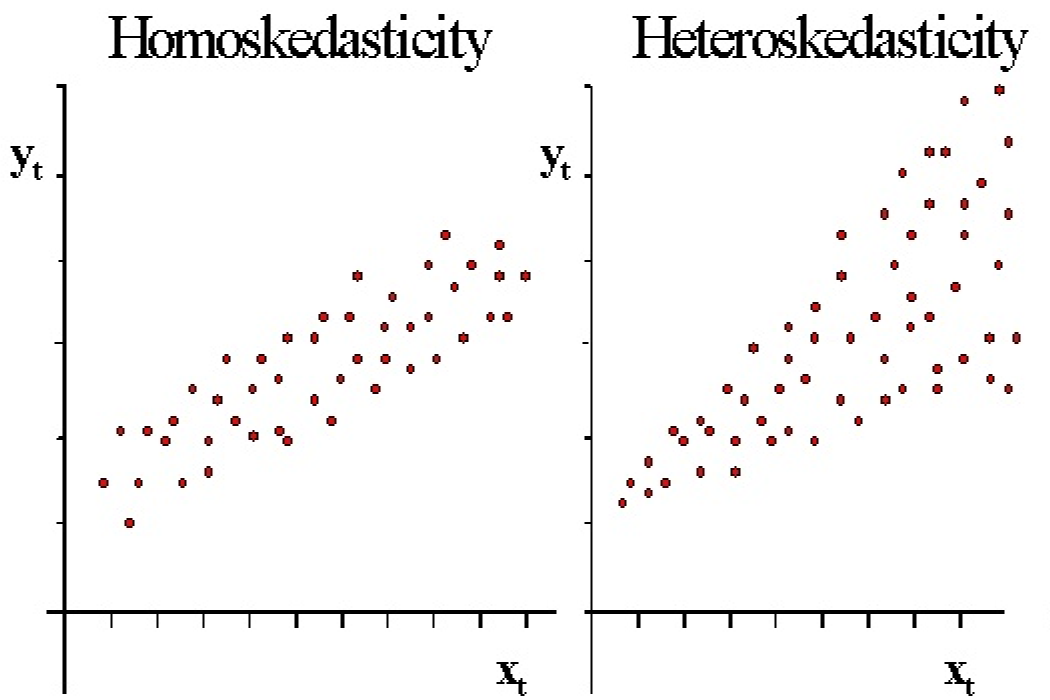
D. A non-scalar identity covariance matrix

1. introduction

Assumption III implies that the covariance matrix of the error vector is a constant σ^2 multiplied by the identity matrix. In general this covariance may be any positive definite matrix. Different assumptions about this matrix will lead to different properties of various estimators.

2. heteroskedasticity

Heteroskedasticity is the case where the diagonal terms of the covariance matrix are not all equal, i.e. Var $(\varepsilon_t) \neq \sigma^2$ for all t



With heteroskedasticity alone the covariance matrix Σ is given by

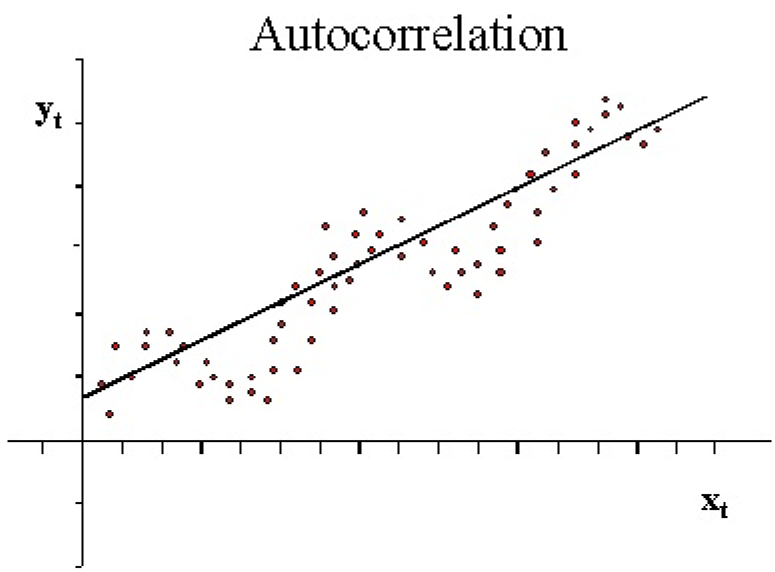
$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \vdots & \sigma_n^2 \end{pmatrix} = \sigma^2 \begin{pmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \omega_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \vdots & \omega_n \end{pmatrix} = \sigma^2 \Omega \tag{22}$$

This model will have k + n parameters and cannot be estimated using n observations unless some assumptions (restrictions) about the parameters are made.

. . .

3. autocorrelation

Autocorrelation is the case where the off-diagonal elements of the covariance matrix are not zero, i.e. Cov (ε_t, ε_s) ≠ 0 for t ≠ s. With no autocorrelation, the errors have no discernible pattern.



In the case above, positive levels of ϵ tend to be associated with positive levels and so on. With autocorrelation alone Σ is given by

$$\Sigma = \begin{pmatrix} \sigma^2 & Cov(\varepsilon_1, \varepsilon_2) & \ldots & Cov(\varepsilon_1, \varepsilon_n) \\ Cov(\varepsilon_2, \varepsilon_1) & \sigma^2 & \ldots & Cov(\varepsilon_2, \varepsilon_n) \\ \vdots & \vdots & \ldots & \vdots \\ Cov(\varepsilon_n, \varepsilon_1) & Cov(\varepsilon_n, \varepsilon_2) & \vdots & \sigma^2 \end{pmatrix} = \sigma^2 \Omega \tag{23}$$

This model will have k + 1 + (n(n-1)/2) parameters and cannot be estimated using n observations unless some assumptions (restrictions) about the parameters are made.

4. the general linear model

For situations in which autocorrelation or heteroskedasticity exists

$$Var(\varepsilon) = \Sigma = \sigma^2\Omega \neq \sigma^2 I \tag{24}$$

and the model can be written more generally as

$$\begin{aligned}
&I \quad y = X\beta + \varepsilon \\
&II \quad E(\varepsilon \mid X) = 0 \\
&III \quad E(\varepsilon\varepsilon' \mid X) = \Sigma = \sigma^2\Omega \\
&IV \quad X \text{ may be fixed or stochastic, but is of rank } k \text{ and generated by a process unrelated to } \varepsilon \\
&V \quad \varepsilon \sim N(0; \Sigma)
\end{aligned} \tag{25}$$

Assumption VI as written here allows X to be stochastic, but along with II, allows all results to be conditioned on X in a meaningful way. This model is referred to as the <u>generalized normal linear regression model</u> and includes the <u>classical normal linear regression model</u> as a special case, i.e., when $\sum = \sigma^2 I$. The unknown parameters in the generalized regression model are the β's = $(\beta_1, \ldots, \beta_k)'$, and the n(n + 1)/2 independent elements of the covariance matrix. In general it is not possible to estimate $\sum$ unless simplifying assumptions are made since one cannot estimate k + [n(n+1)/2] parameters with n observations.

5. Least squares estimations of β in the general linear model with $\Sigma = \sigma^2\Omega$ known

Least squares estimation makes no assumptions about the disturbance matrix and so is defined as before using the sum of squared errors. The sum of squared errors is defined by

$$\begin{aligned}
SSE(\hat{\beta}) &= \sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n} (y_t - x_t'\beta)^2 \\
&= \begin{bmatrix} e_1 & e_2 & . & . & . & e_n \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \\
&= e'e \\
&= (y - X\hat{\beta})'(y - X\hat{\beta}) \\
&= y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\
&= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}
\end{aligned} \tag{26}$$

The necessary condition for $SSE(\hat{\beta})$ to be a minimum is that

$$\begin{aligned}
&\frac{\partial SSE(\hat{\beta})}{\partial \hat{\beta}'} = 0 \\
&\Rightarrow -2X'y + 2X'X\hat{\beta} = 0
\end{aligned} \tag{27}$$

This gives the normal equations which can then be solved to obtain the least squares estimator

$$\begin{aligned} X'X\hat{\beta} &= X'y \\ \Rightarrow \hat{\beta} &= (X'X)^{-1}X'y \end{aligned} \qquad \textbf{(28)}$$

The least squares estimator is exactly the same as before. Its properties may be different, however, as will be shown in a later section.

6. Maximum likelihood estimation with Σ known

The likelihood function for the vector random variable y is given by the multivariate normal density. For this model

$$y \sim N(X\beta;\, \Sigma). \qquad \textbf{(29)}$$

Therefore the likelihood function is given by

$$L(y;\, \beta,\, \Sigma) = \frac{e^{\frac{-1}{2}(y - X\beta)'\,\Sigma^{-1}\,(y - X\beta)}}{(2\pi)^{\frac{n}{2}}\,|\Sigma|^{\frac{1}{2}}} \qquad \textbf{(30)}$$

The natural log of the likelihood function is given as

$$\begin{aligned} \ell = \log L &= \frac{-1}{2}\,(y - X\beta)'\,\Sigma^{-1}\,(y - X\beta) - \frac{n}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| \\ &= \frac{-1}{2}\left[y'\Sigma^{-1}y - 2\beta'X'\Sigma^{-1}y + \beta'X'\Sigma^{-1}X\beta\right] - \frac{n}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| \end{aligned} \qquad \textbf{(31)}$$

The M.L.E. of β is defined by maximizing 31

$$\frac{\partial \ell}{\partial \beta'} = -\frac{1}{2}\left(-2X'\Sigma^{-1}y + 2X'\Sigma^{-1}X\beta^{\ell}\right) = 0 \qquad \textbf{(32)}$$

This then yields as an estimator of β

$$\beta^{\ell} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y \qquad \textbf{(33)}$$

This estimator differs from the least squares estimator. Thus the least squares estimator will have different properties than the maximum likelihood estimator. Notice that if Σ is equal to $\sigma^2 I$, the estimators are the same.

$$\begin{aligned}
\beta^{\ell} &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y \\
&= (X'(\sigma^2 I)^{-1}X)^{-1}X'(\sigma^2 I)^{-1}y \\
&= (X'(\sigma^2)^{-1}X)^{-1}X'(\sigma^2)^{-1}y \\
&= \left(X'\left(\frac{1}{\sigma^2}\right)X\right)^{-1}X'\left(\frac{1}{\sigma^2}\right)y \\
&= \sigma^2(X'X)^{-1}X'\left(\frac{1}{\sigma^2}\right)y \\
&= (X'X)^{-1}X'y
\end{aligned} \tag{34}$$

7. Best linear unbiased estimation with Σ known

BLUE estimators are obtained by finding the best estimator that satisfies certain conditions. BLUE estimators have the properties of being linear, unbiased, and minimum variance among all linear unbiased estimators. Linearity and unbiasedness can be summarized as

$$\begin{aligned}
\tilde{\beta} &= Ay \quad A \text{ is } k \times n \quad \textit{linearity} \\
E(\tilde{\beta}) &= E(Ay) = AE(y) = AX\beta = \beta \\
\Longleftrightarrow AX &= I \quad \textit{unbiasedness}
\end{aligned} \tag{35}$$

The estimator must also be minimum variance. One definition of this is that the variance of each $\tilde{\beta}_i$ must be a minimum. The variance of the ith β is given by the ith diagonal element of

$$\begin{aligned}
Var(\tilde{\beta}) &= A\,Var(y)A' \\
&= A\Sigma A'
\end{aligned} \tag{36}$$

This can be denoted as

$$\begin{aligned}
Var(\tilde{\beta}_i) &= a_i'\,Var(y)\,a_i \\
&= a_i'\,\Sigma\,a_i
\end{aligned} \tag{37}$$

where a_i' is the ith row of the matrix A, $\tilde{\beta}_i$ is $a_i'y$ and a_i' is given as $a_i' = i'A$ where i′ is the ith row of an kxk identity matrix. The construction of the estimator can be reduced to selecting the matrix A so that the rows of A

$$\begin{aligned}
&\min\; a_i'\,\Sigma\,a_i \quad i = 1,2,\ldots,k \\
&s.t. \quad AX = I
\end{aligned} \tag{38}$$

Because the result will be symmetric for each β_i (hence, for each a_i), denote a_i' by a' where a is an (n by 1) vector. The problem then becomes:

$$
\begin{aligned}
&\min\ a'\Sigma a \qquad \Sigma\ is\ n \times n \\
&s.t.\ AX = I \quad X\ is\ n \times k \\
&or \\
&\min\ a'\Sigma a \\
&s.t.\ X'a = i
\end{aligned} \tag{39}
$$

The column vector i is the ith column of the identity matrix. The Lagrangian is as follows

$$
\ell = a'\Sigma a - \lambda'(X'a - i) \qquad \lambda\ is\ k \times 1 \tag{40}
$$

To minimize it take the derivatives with respect to a and λ

$$
\begin{aligned}
\frac{\partial \ell}{\partial a} &= 2a'\Sigma - \lambda'X' = 0 \\
\frac{\partial \ell}{\partial \lambda'} &= -(X'a - i) = 0 \\
\Rightarrow a' &= \frac{1}{2}\lambda'X'\Sigma^{-1} \\
and\ a'X &= i'
\end{aligned} \tag{41}
$$

Now substitute $a' = (1/2)\lambda'X'\Sigma^{-1}$ into the second equation in 41 to obtain

$$
\begin{aligned}
&\frac{1}{2}\lambda'X'\Sigma^{-1}X - i' = 0 \\
\Rightarrow \lambda' &= 2i'(X'\Sigma^{-1}X)^{-1} \\
\Rightarrow a' &= \frac{1}{2}2i'(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1} \\
&= i'(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1} \\
\Rightarrow A &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1} \\
\Rightarrow \tilde{\beta} &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y
\end{aligned} \tag{42}
$$

It is obvious that AX = I.

The BLUE and MLE estimators of β are identical, but different from the least squares estimator of β. We sometimes call the BLUE estimator of β in the general linear model, the generalized least squares estimator, β_{GLS}. This estimator is also sometimes called the Aitken estimator after the individual who first proposed it.

8. A note on the distribution of $\hat{\beta}$, β^{ℓ}, *and* $\tilde{\beta}$

a. introduction

For the Classical Normal Linear Regression Model we showed that

$$\begin{aligned} \varepsilon &\sim N(0,\ \sigma^2 I) \\ \hat{\beta} &= \tilde{\beta} = \beta^{\ell} = (X'X)^{-1}X'y \\ \hat{\beta} &\sim N\left(\beta;\sigma^2(X'X)^{-1}\right) \end{aligned} \tag{43}$$

For the Generalized Regression Model

$$\begin{aligned} \varepsilon &\sim N(0,\Sigma) \\ \hat{\beta} &= (X'X)^{-1}X'y = A_1 y \\ \tilde{\beta} &= \beta^{\ell} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y = A_2 y \end{aligned} \tag{44}$$

b. unbiasedness of ordinary least squares in the general linear model

As before write $\hat{\beta}$ in the following fashion.

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'y \\ &= (X'X)^{-1}X'(X\beta + \epsilon) \\ &= \beta + (X'X)^{-1}X'\varepsilon \end{aligned} \tag{45}$$

Now take the expected value of equation 45

$$\begin{aligned} E[\hat{\beta} \mid X) &= \beta + E[(X'X)^{-1}X'\varepsilon \mid X] \\ &= \beta + (X'X)^{-1}X' E(\varepsilon \mid X) \\ &= \beta \end{aligned} \tag{46}$$

Because $(X'X)^{-1}X'$ is either fixed or a function only of X if X is stochastic, it can be factored out of the expectation, leaving E(ϵ|X), which has an expectation of zero by assumption II. Now find the unconditional expectation of $\hat{\beta}$ by using the law of iterated expectations. In the sense of Theorem 3 of the section on alternative estimators, h(X,Y) is $\hat{\beta}$ and $E_{Y|X}$ computes the expected value of $\hat{\beta}$ conditioned on X.

$$E[\hat{\beta}] = E_X\left[E(\hat{\beta} \mid X)\right] = E_X[\beta] = \beta \tag{47}$$

The interpretation of this result is that for any particular set of observations, X, the least squares estimator has expectation β.

c. variance of the OLS estimator

First rewrite equation 42 as follows

$$\hat{\beta} = \beta + (X'X)^{-1}X'\varepsilon$$
$$\Rightarrow \hat{\beta} - \beta = (X'X)^{-1}X'\varepsilon \tag{48}$$

Now directly compute the variance of $\hat{\beta}$ given X.

$$
\begin{aligned}
Var(\hat{\beta} \mid X) &= E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)' \mid X\right] \\
E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)' \mid X\right] &= E\left[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1} \mid X\right] \\
&= (X'X)^{-1}X'E(\epsilon\epsilon' \mid X)\,X(X'X)^{-1} \\
&= (X'X)^{-1}X'\sigma^2\,\Omega\,X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}X'\,\Omega\,X(X'X)^{-1} \\
&= \frac{\sigma^2}{n}\left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{n}X'\,\Omega\,X\right)\left(\frac{1}{n}X'X\right)^{-1}
\end{aligned}
\tag{49}
$$

If the regressors are non-stochastic, then this is also the unconditional variance of $\hat{\beta}$. If the regressors are stochastic, then the unconditional variance is given by $E_X\left[Var\left(\hat{\beta} \mid X\right)\right]$

d. unbiasedness of MLE and BLUE in the general linear model

First write the GLS estimator as follows

$$
\begin{aligned}
\tilde{\beta} &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y \\
&= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}(X\beta + \epsilon) \\
&= \beta + (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\varepsilon
\end{aligned}
\tag{50}
$$

Now take the expected value of equation 50

$$
\begin{aligned}
E[\tilde{\beta} \mid X) &= \beta + E[(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\varepsilon \mid X] \\
&= \beta + (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}E(\varepsilon \mid X) \\
&= \beta
\end{aligned}
\tag{51}
$$

Because $(X'X)^{-1}X'$ is either fixed or a function only of X if X is stochastic, it can be factored out of the expectation, leaving E(ϵ|X), which has an expectation of zero by assumption II. Now find the unconditional expectation of $\tilde{\beta}$ by using the law of iterated expectations.

$$E[\tilde{\beta}] = E_X\left[E(\tilde{\beta} \mid X)\right] = E_X\left[\beta\right] = \beta \tag{52}$$

The interpretation of this result is that for any particular set of observations, X, the generalized least squares estimator has expectation β.

e. variance of the GLS (MLE and BLUE) estimator

First rewrite equation 50 as follows

$$\begin{aligned} \tilde{\beta} &= \beta + (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\varepsilon \\ \Rightarrow \tilde{\beta} - \beta &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\varepsilon \end{aligned} \tag{53}$$

Now directly compute the variance of $\tilde{\beta}$ given X.

$$\begin{aligned} Var(\tilde{\beta} \mid X) &= E\left[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)' \mid X\right] \\ E\left[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)' \mid X\right] &= E\left[(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\varepsilon\varepsilon'\Sigma^{-1}X(X'\Sigma^{-1}X)^{-1} \mid X\right] \\ &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}E(\epsilon\epsilon' \mid X)\,\Sigma^{-1}X(X'X)^{-1} \\ &= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\,\Sigma\,\Sigma^{-1}X(X'\Sigma^{-1}X)^{-1} \\ &= (X'\Sigma^{-1}X)^{-1} \\ &= \sigma^2(X'\Omega^{-1}X)^{-1} \\ &= \frac{\sigma^2}{n}\left(\frac{1}{n}X'\Omega^{-1}X\right)^{-1} \end{aligned} \tag{54}$$

If the regressors are non-stochastic, then this is also the unconditional variance of $\tilde{\beta}$. If the regressors are stochastic, then the unconditional variance is given by $E_X\left[Var(\tilde{\beta} \mid X)\right]$

f. summary of finite sample properties of OLS in the general model

Note that all the estimators are unbiased estimators of β, but $Var(\hat{\beta}) \neq Var(\beta^{\ell}) = Var(\tilde{\beta})$. If $\Sigma = \sigma^2 I$ then the classical model results are obtained. Thus using least squares in the generalized model gives unbiased estimators, but the variance of the estimator may not be minimal.

9. Consistency of OLS in the generalized linear regression model

We have shown that the least squares estimator in the general model is unbiased. If we can show that its variance goes to zero as n goes to infinity we will have shown that it is mean square error consistent, and thus that is converges in probability to β. This variance is given by

$$\begin{aligned} Var(\hat{\beta} \mid X) &= \sigma^2(X'X)^{-1}X'\,\Omega\,X(X'X)^{-1} \\ &= \frac{\sigma^2}{n}\left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{n}X'\,\Omega\,X\right)\left(\frac{1}{n}X'X\right)^{-1} \end{aligned} \tag{55}$$

As previously, we will assume that

$$\begin{aligned} &X \text{ is of rank } k \\ &\text{and } \operatorname*{plim}_{n\to\infty} \frac{X'X}{n} = Q \\ &\text{where } Q \text{ exists as a finite nonsingular matrix} \end{aligned} \tag{56}$$

With this assumption, we need to consider the remaining term, i.e.,

$$\frac{\sigma^2}{n}\left(\frac{1}{n}X'\,\Omega\,X\right) \tag{57}$$

The leading term, $\frac{\sigma^2}{n}$, will, by itself go to zero. We can write the matrix term in the following useful fashion similar to the way we wrote out a matrix product in proving the asymptotic normality of the non-linear least squares estimator. Remember specifically that

$$X'X = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1k} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & x_{3n} & \cdots & x_{nk} \end{pmatrix} = \begin{pmatrix} \sum_{t=1}^{n} x_{t1}^2 & \sum_{t=1}^{n} x_{t1}x_{t2} & \sum_{t=1}^{n} x_{t1}x_{t3} & \cdots & \sum_{t=1}^{n} x_{t1}x_{tk} \\ \sum_{t=1}^{n} x_{t2}x_{t1} & \sum_{t=1}^{n} x_{t2}^2 & \sum_{t=1}^{n} x_{t2}x_{t3} & \cdots & \sum_{t=1}^{n} x_{t2}x_{tk} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{t=1}^{n} x_{tk}x_{t1} & \sum_{t=1}^{n} x_{tk}x_{t2} & \sum_{t=1}^{n} x_{tk}x_{t3} & \cdots & \sum_{t=1}^{n} x_{tk}^2 \end{pmatrix} = \sum_{t=1}^{n} x_{t.}'\,x_{t.} \tag{58}$$

where In similar fashion we can show that the matrix in equation 57 can be written as

$$\frac{\sigma^2}{n}\left(\frac{1}{n}X'\,\Omega\,X\right) = \frac{\sigma^2}{n}\left(\frac{\sum_{t=1}^{n}\sum_{\tau=1}^{n} \omega_{t\tau}\, x_{t.}'\, x_{\tau.}}{n}\right) \tag{59}$$

The second term in equation 59 is a sum of n^2 terms divided by n. In order to check convergence of this product, we need to consider the order of each term. Remember the definition of order given earlier.

Definition of order:

1. A sequence $\{a_n\}$ is at most of order n^λ, which we denote $O(n^\lambda)$ if $n^{-\lambda}a_n = \frac{a_n}{n^\lambda}$ is bounded. When $\lambda = 0$, $\{a_n\}$ is bounded, and we also write $a_n = O(1)$, which we say as *big oh one*.
2. A sequence $\{a_n\}$ is of smaller order then n^λ, which we denote $o(n^\lambda)$ if $n^{-\lambda}a_n = \frac{a_n}{n^\lambda} \to 0$. When $\lambda = 0$, $\{a_n\}$ converges to zero, and we also write $a_n = o(1)$, which we say as *little oh one*.

The first term in the product is of order 1/n, O(1/n). The second term, in general is of O(n). So it appears that if the product of these two terms converges, it might converge to a matrix of non-zero constants. If this were the case, proving consistency would be a problem. At this point we will simply make an assumption as follows.

$$\underset{n\to\infty}{plim} \frac{X'\Omega X}{n} \text{ exists as a finite nonsingular matrix} \tag{60}$$

If this is the case, then the expression in equation 59 will converge in the limit to zero, and $\hat{\beta}$ will be consistent. Using arguments similar to those adopted previously, we can also show that the OLS estimator will be asymptotically normal in a wide variety of settings (Amemiya, 187). Discussion of the GLS estimator will be discussed later.

10. Consistent estimators of the covariance matrix in the case of general error structures

a. general discussion

If Ω were known, then the estimator of the asymptotic covariance matrix of $\hat{\beta}$ would be

$$\begin{aligned} Var(\hat{\beta} \mid X) &= \sigma^2 (X'X)^{-1} X' \Omega X (X'X)^{-1} \\ &= \frac{\sigma^2}{n}\left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{n}X'\Omega X\right)\left(\frac{1}{n}X'X\right)^{-1} \\ &= \frac{1}{n}\left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{n}X'\sigma^2\Omega X\right)\left(\frac{1}{n}X'X\right)^{-1} \end{aligned} \tag{61}$$

The outer terms are available from the data, and if $\sigma^2\Omega$ were known, we would have the information we need to compute standard errors. From a sample of n observations, there is no way to estimate the elements of $\sigma^2\Omega$. But what we really need is an estimator of $\underset{n\to\infty}{plim} \frac{X'\sigma^2\Omega X}{n}$ which is a symmetric k×k matrix. What we then need is an estimator of this matrix. We can write this is a more useful fashion as follows

$$\underset{n\to\infty}{plim} Q_* = \underset{n\to\infty}{plim} \frac{1}{n}\left(X'\sigma^2\Omega X\right) = \frac{1}{n}\left(\sum_{t=1}^{n}\sum_{\tau=1}^{n} \sigma_{t\tau} x_{t.}' x_{\tau.}\right) \tag{62}$$

where $\sigma_{t\tau}$ is the appropriate element of $\sigma^2\Omega$ as compared to Ω. The idea will be to use information on the residuals from the least squares regression to devise a way to approximate Q_*.

b. heteroskedasticity only

In the case where there is no auto correlation, that is when $\sigma^2\Omega$ is a diagonal matrix, we can write equation 62 as

$$\underset{n\to\infty}{plim} Q_* = \underset{n\to\infty}{plim} \frac{1}{n}\left(X'\sigma^2\Omega X\right) = \frac{1}{n}\left(\sum_{t=1}^{n} \sigma_t^2 x_{t.}' x_{t.}\right) \tag{63}$$

White has shown that under very general conditions, the estimator

$$S_0 = \frac{1}{n}\left(\sum_{t=1}^{n} e_t^2 x_{t.}' x_{t.}\right) \overset{P}{\rightarrow} plim\, Q_* \tag{64}$$

The proof is based on the fact that $\hat{\beta}$ is a consistent estimate of β, (meaning the residuals are consistent estimates of ε), and fairly mild assumptions on X. Then rather than using $\sigma^2(X'X)^{-1}$ to estimated the variance of $\hat{\beta}$ in the general model, we instead use

$$\begin{aligned} Est\left(Var(\hat{\beta} \mid X)\right) &= \frac{1}{n}\left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{n}\sum_{t=1}^{n} e_t^2 x_{t.}' x_{\tau.}\right)\left(\frac{1}{n}X'X\right)^{-1} \\ &= n(X'X)^{-1} S_0 (X'X)^{-1} \end{aligned} \tag{65}$$

c. autocorrelation

In the case of a more general covariance matrix, a candidate estimator for Q_* might be

$$Q_* = \frac{1}{n}\left(\sum_{t=1}^{n}\sum_{\tau=1}^{n} e_t e_\tau x_{t.}' x_{\tau.}\right) \tag{66}$$

The difficulty here is that to this matrix may not converge in the limit. To obtain convergence, it is necessary to assume that the terms involving unequal subscripts in (66) diminish in importance as n grows. A sufficient condition is that terms with subscript pairs $|t - \tau|$ grow smaller as the distance between then grows larger. A more practical problem for estimation is that Q_* may not be positive definite. Newey and West have proposed an estimator to solve this problem using some of the cross products $e_t e_{t-\ell}$. This estimator will be discussed in a later section.

E. Stochastic X matrix (possibly less than full rank)

1. X matrix less than full rank

If the X matrix, which is nxk, has rank less than k then X'X cannot be inverted and the least squares estimator will not be defined. This was discussed in detail in the section on multicollinearity.

2. Stochastic X

Consider the least squares estimator of β in the classical model. We showed that it was unbiased as follows.

$$\begin{aligned} E[\hat{\beta} \mid X] &= \beta + E[(X'X)^{-1}X'\varepsilon \mid X] \\ &= \beta + (X'X)^{-1}X'\, E(\varepsilon \mid X) \\ &= \beta \end{aligned} \tag{67}$$

If the X matrix is stochastic and correlated with ϵ, we cannot factor it out of the second term in equation 48. If this is the case, $E(\hat{\beta}) \neq \beta$. In such cases, the least squares estimator is usually not only biased, but is inconsistent as well. Consider for example the case where $\left(\frac{X'X}{n}\right)$ converges to a finite and non-singular matrix Q. Then we can compute the probability limit of $\hat{\beta}$ as

$$\begin{aligned} plim\ \hat{\beta} &= \beta + plim\left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'\varepsilon}{n}\right) \\ &= \beta + Q^{-1} plim\left[\frac{X'\varepsilon}{n}\right] \\ &\neq \beta \end{aligned} \tag{68}$$

unless $plim\left[\frac{X'\varepsilon}{n}\right] = 0$.

We showed previously that a consistent estimator of β could be obtained using instrumental variables (IV). The idea of instrumental variables is to devise an estimator of β such that the second term in equation 49 will have a probability limit of zero. The instrumental variables estimator is based on the idea that the "instruments" used in the estimation are not highly correlated with ε and any correlation disappears in large samples. A further condition is that these instruments are correlated with variables in the matrix X. We defined instrumental variables estimators in two different cases, when the number of instrumental variables was equal to the number of columns of the X matrix, i.e., Z was n x k matrix, and cases where there were more than k instruments. In either case we assumed that had the following properties

$$\begin{aligned} &plim\ \frac{Z'X}{n} = \Sigma_{ZX} \quad \text{is finite and nonsingular} \\ &plim\ \frac{Z'\varepsilon}{n} = 0 \\ &plim\ \frac{Z'y}{n} = \Sigma_{Zy} \quad \text{exists} \end{aligned} \tag{69}$$

Then the instrumental variables estimator was given by

$$\beta_{IV} = (Z'X)^{-1} Z'y \tag{70}$$

By finding the plim of this estimator, we showed that it was consistent

$$\begin{aligned} plim(\beta_{IV}) &= plim[(Z'X)^{-1} Z'y] \\ &= plim[(Z'X)^{-1} Z'X\beta + (Z'X)^{-1} Z'\varepsilon] \\ &= \beta + plim\left[\left(\frac{Z'X}{n}\right)^{-1} \frac{Z'\epsilon}{n}\right] \\ &= \beta + plim\left(\frac{Z'X}{n}\right)^{-1} plim\left(\frac{Z'\epsilon}{n}\right) \\ &= \beta \end{aligned} \tag{71}$$

In the case where the number of instrumental variables was greater than k, we formed k instruments by projecting each of the columns of the stochastic X matrix on all of the instruments, and then used the predicted values of X from these regressions as instrumental variables in defining the IV estimator. If we let P_Z be the matrix that projects orthogonally onto the column space defined by the vectors Z, S(Z), then the IV estimator is given by

$$\begin{aligned} \beta_{IV} &= (X' P_Z X)^{-1} X' P_Z y \\ &= (X' Z(Z' Z)^{-1} Z' X)^{-1} X' Z(Z' Z)^{-1} Z' y. \end{aligned} \tag{72}$$

We always assume that the matrix $X'P_ZX$ has full rank, which is a necessary condition for β_{IV} to be identified. In a similar fashion to equation 52, we can show that this IV estimator is consistent.

F. Random disturbances are not distributed normally (assumptions I-IV hold)

1. General discussion

An inspection of the derivation of the least squares estimator $\hat{\beta}$ reveals that the deduction is not dependent upon any of the assumptions II-V except for the full rank condition on X. It really doesn't depend on I, if we are simply estimating a linear model no matter the nature of the underlying model. Thus for the model

$$y = X\beta + \varepsilon \tag{73}$$

the OLS estimator is always

$$\hat{\beta} = (X'X)^{-1} X'y \tag{74}$$

even when assumption V is dropped. However, the statistical properties of $\hat{\beta}$ are very sensitive to the distribution of ε.

Similarly, we note that while the BLUE of β depends upon II-IV, $\hat{\beta}$ is invariant with respect to the assumptions about the underlying density of ε as long as II-IV are valid. We can thus conclude that

$$\hat{\beta} = \tilde{\beta} = (X'X)^{-1} X'y \tag{75}$$

even when the error term is not normally distributed.

2. Properties of the estimators (OLS and BLUE) when ε is not normally distributed

When the error terms in the linear regression model are not normally distributed, the OLS and BLUE estimators are:

a. unbiased
b. minimum variance of all unbiased linear estimators (not necessarily of all unbiased estimators since the Cramer Rao lower bound is not known unless we know the density of the residuals)
c. consistent
d. but standard t and F tests and confidence intervals are not necessarily valid for nonnormally

distributed residuals

The distribution of $\hat{\beta}$ (e.g., normal, Beta, Chi Square, etc.) will depend on the distribution of ε which determines the distribution of y ($y = X\beta + \varepsilon$).

The maximum likelihood estimator, of course, will differ since it depends explicitly on the joint density function of the residuals. And this joint density gives rise to the likelihood function

$$L = f(y_1;\beta)\, f(y_2;\beta) \cdots f(y_n;\beta) \tag{76}$$

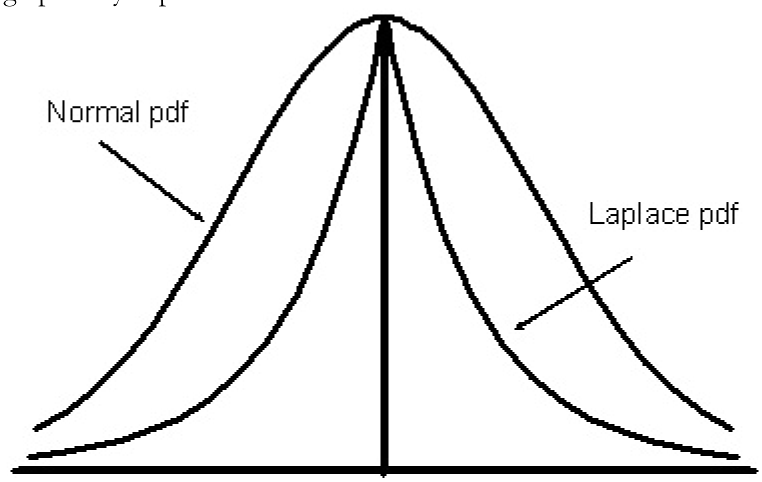
and requires a knowledge of the distribution of the random disturbances. It is not defined otherwise. MLE are generally efficient estimators and least squares estimators will be efficient if f(y;·) is normal. However, least squares need not be efficient if the residuals are not distributed normally.

3. example

Consider the case in which the density function of the random disturbances is defined by the Laplace distribution

$$f(\varepsilon_t) = \frac{1}{2} e^{-|\varepsilon_t|} \qquad -\infty < \varepsilon_t < \infty \tag{77}$$

which can be graphically depicted as



The associated likelihood function is defined by

$$\begin{aligned} L &= f(y_1; \beta)\, f(y_2; \beta) \cdots f(y_n; \beta) \\ &= \frac{e^{-|y_1 - x_1\beta|}}{2} \quad . \; . \; . \quad \frac{e^{-|y_n - x_n\beta|}}{2} \end{aligned} \tag{78}$$

where $x_t = (1, x_{t2}, \ldots, x_{tk})$, $\beta' = (\beta_1, \ldots, \beta_k)$. The log likelihood function is given by

$$\ell = \log L = -\Sigma \; |y_t - x_t\beta \;| - n\log(2) \tag{79}$$

The MLE of β in this case will minimize

$$\sum_{t=1}^{n} |y_t - x_t\beta \;| \tag{80}$$

and is sometimes called the "least lines," minimum absolute deviations (MAD), or least absolute deviation (LAD) estimator. It will have all the properties of maximum likelihood estimators such as being asymptotically unbiased, consistent, and asymptotically efficient. It need not, however, be unbiased, linear, or minimum variance of all unbiased estimators.

4. Testing for and using other distributions

The functional form of the distribution of the residuals is rarely investigated. This can be done, however, by comparing the distribution of ε_t with the normal.

Various tests have been proposed to test the assumption of normality. These tests take different forms. One class of tests is based on examining the skewness or kurtosis of the distribution of the estimated residuals. Chi square goodness of fit tests have been proposed which are based upon comparing the histogram of estimated residuals with the normal distribution. The Kolmogorov-Smirnov test is based upon the distribution of the maximum vertical distance between the cumulative histogram and the cumulative distribution of the hypothesized distribution.

An alternative approach is to consider general distribution functions such as the beta or gamma which include many of the common alternative specifications as special cases.

Literature Cited

Aitken, A.C. "On Least Squares and Liner Combinations of Observations." *Proceedings of the Royal Statistical Society*, 55, (1935):42-48

Amemiya, T. *Advanced Econometrics*. Cambridge: Harvard University Press, 1985.

Huber, P. J., *Robust Statistics*, New York: Wiley, 1981.

Newey, W., and K. West. "A Simple Positive Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica* 55 (1987): 703-708

White, H. "A Heteroskedasticity Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48 (1980): 817-838.