# Economics 620, Lecture 7: Still More, But Last, on the K-Varable Linear Model

Nicholas M. Kiefer

Cornell University

# Specification Error:

Suppose the model generating the data is

$$y = X\beta + \varepsilon$$

However, the model fitted is $y = X^*\beta^* + \varepsilon$, with the LS estimator

$$b^* = (X^{*\prime}X^*)^{-1}X^{*\prime}y$$
$$= (X^{*\prime}X^*)^{-1}X^{*\prime}X\beta + (X^{*\prime}X^*)^{-1}X^{*\prime}\varepsilon.$$

Then $Eb^* = (X^{*\prime}X^*)^{-1}X^{*\prime}X\beta$ and $V(b^*) = \sigma^2(X^{*\prime}X^*)^{-1}$

# Application 1: Excluded variables

Let $X = [X_1 X_2]$ and $X^* = X_1$.
That is, the model that generates the data is

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon.$$

Consider $b^*$ as an estimator of $\beta_1$.

*Proposition*: $b^*$ is biased.

*Proof*:

$$\begin{aligned}
b^* &= (X^{*\prime}X^*)^{-1}X^{*\prime}y \\
&= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + \varepsilon) \\
&= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\varepsilon \\
Eb^* &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2
\end{aligned}$$

The second expression on the right hand side is the bias. ■

# A classic example:

Suppose that the model generating the data is
$y_i = \beta_0 + \beta_1 S_i + a_i + \varepsilon_i$
$y$: natural logarithm of earnings
$S$: schooling
$a$: ability

$a$ is unobserved and omitted, but it is positively correlated with $S$.
Then

$$Eb^* = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} N & \sum S \\ \sum S & \sum S^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum a \\ \sum aS \end{bmatrix}$$

supposing $a$ is measured so that its coefficient is 1.

If we suppose that $\sum a = 0$, then the bias in the coefficient of schooling is positive.

# A classic example (cont'd)

Generally, we cannot sign the bias, it depends not only on $\beta_2$ but also on $(X_1'X_1)^{-1}X_1'X_2$, which of course can be positive or negative.

Note that $Vb^* = \sigma^2(X_1'X_1)^{-1}$. So if $\beta_2 = 0$, there is an efficiency gain from imposing the restriction and leaving out $X_2$. This confirms our earlier results.

# Estimation of Variance:

$e^* = M_1 y = M_1(X_1\beta_1 + X_2\beta_2 + \varepsilon)$
$= M_1 X_2 \beta_2 + M_1 \epsilon$
$\Rightarrow e^{*\prime} e^* = \beta_2' X_2' M_1 X_2 \beta_2 + \varepsilon' M_1 \varepsilon + 2\beta_2' X_2' M_1 \epsilon$
Note the expected value of the last term is 0.

Clearly, we cannot estimate $\sigma^2$ by usual methods even if $X_1' X_2 = 0$ (no bias) since still $M_1 X_2 \neq 0$.

There is hope of detecting misspecification from the residuals since
$Ee^* e^{*\prime} = \sigma^2 M_1$ under correct specification and
$Ee^* e^{*\prime} = \sigma^2 M_1 + M_1 X_2 \beta_2 \beta_2' X_2' M_1$ under misspecification.

# Application 2:

Inclusion of unnecessary variables.

Let $X = X_1$ and $X^* = [X_1 \quad X_2]$
$X_1$ is $N \times K_1$ and $X_2$ is $N \times K_2$.
That is, the "true" model is $y = X_1\beta + \varepsilon$.

*Proposition*: $b^*$ is unbiased.

*Proof*: $$Eb^* = (X^{*\prime}X^*)^{-1}X^{*\prime}X\beta$$
$$= \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1} \begin{bmatrix} X_1'X_1 \\ X_2'X_1 \end{bmatrix}\beta$$

The partitioned inversion formula gives

$$\left[\begin{array}{cc} \Gamma & -(X_1'X_1)^{-1}X_1'X_2D \\ -DX_2'X_1(X_1'X_1)^{-1} & D \end{array}\right]$$

for $(X^{*\prime}X^*)^{-1}$ where $D = (X_2'M_1X_2)^{-1}$ and

$$\Gamma = (X_1'X_1)^{-1} + (X_1'X_1)^{-1}X_1'X_2DX_2'X_1(X_1'X_1)^{-1}.$$

This is a symmetric matrix. Multiplying this out verifies that

$$Eb^* = \left[\begin{array}{c} \beta \\ 0 \end{array}\right]. \blacksquare$$

Note that the variance of $b^*$ is

$$V(b^*) = \sigma^2(X^{*\prime}X^*)^{-1}.$$

*Proposition*: The variance of the coefficients of $X_1$ in the unrestricted (where the matrix of explanatory variables is $X^*$) is greater than the variance of the coefficients in the restricted model (where the matrix of explanatory variables is $X_1$).

*Proof*: Using partitioned inversion, the variance of the first $K_1$ elements (coefficients on $X_1$) is $\sigma^2(X_1'M_2X_1)^{-1} \geq \sigma^2(X_1'X_1)^{-1} =$ variance of the restricted estimator. (*why*?) ■

*Estimation of* $\sigma^2$:
$e^* = M^*y = M^*\varepsilon$

Under normality,
$(e^{*\prime}e^*/\sigma^2) = (\varepsilon' M^*\varepsilon/\sigma^2) \sim \chi^2(N - K_1 - K_2)$

$\Rightarrow s^2$ has higher variance than in the restricted model. (*why*?)

# Note on the interpretation of bias:

$Ey = X\beta$ *defines* $\beta$ and $LS$ gives unbiased estimates of *that* $\beta$. Questions of bias really require a model.

Further statistical assumptions like

$$Vy = \sigma^2 I$$

allow some sorting out of specifications, but is this assumption really attractive?

# Cross products matrix:

In $LS$, "everything" comes from the cross products matrix.

*Definition*: The cross products matrix is

$$\begin{bmatrix} y'y & y'X \\ X'y & X'X \end{bmatrix} = \begin{bmatrix} \sum y_i^2 & \sum y_i & \sum y_i x_{2i} & \cdots \sum y_i x_{Ki} \\ \bullet & \sum 1 & \sum x_{2i} & \cdots \sum x_{Ki} \\ \bullet & \bullet & \sum x_{2i}^2 & \cdots \sum x_{2i} x_{Ki} \\ \bullet & \bullet & \bullet & \cdots \sum x_{Ki}^2 \end{bmatrix}$$

with a column of ones in $X$.
It is a symmetric matrix.
Note that $x_j$ refers to the $j$th column of $X$.

# Heteroskedasticity

$V(Y) = V \neq \sigma^2 I$

Is the $LS$ estimator unbiased? Is it BLUE?

*Proposition*: Under the assumption of heteroskedasticity, $V(\hat{\beta}) = (X'X)^{-1}X'VX(X'X)^{-1}$.

*Proof*:
$V(\hat{\beta}) = E(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}$
$= (X'X)^{-1}X'VX(X'X)^{-1}$. ∎

Suppose $\sum = V(\varepsilon)$ is a diagonal matrix. Then

$$X' \sum X = E \sum X_i \varepsilon_i^2 X_i'.$$

Note that the cross products have expectation 0.
This suggests using $\sum X_i e_i^2 X_i'$.
So we can estimate standard errors under the assumption that $V(y)$ is diagonal.

# Testing for heteroskedasticity:

**1. Goldfeld-Quandt test:**

Suppose we suspect that $\sigma_i^2$ varies with $x_i$. Then reorder the observations in the order of $x_i$. Suppose $N$ is even. If $\varepsilon$ was observed, then

$$\frac{\varepsilon_1^2 + \varepsilon_2^2 + \ldots + \varepsilon_{N/2}^2}{\varepsilon_{[(n/2)+1]}^2 + \varepsilon_{[(N/2)+2]}^2 + \ldots + \varepsilon_N^2} \sim F(N/2, N/2)$$

could be used.

We are tempted to use $e_i$, but we can't because the first $N/2$ $e_i$'s are not independent of the last.

Here comes the Goldfeld-Quandt trick: Estimate $e$ separately for each half of the sample with $K$ parameters. The statistic is $F((N/2) - K,$ $(N/2) - K)$.
It turns out that this "works" better if you delete the middle $N/3$ observations.

# Testing for heteroskedasticity (cont'd):

**2. Breusch-Pagan test:**

The disturbances $\varepsilon_i$ are assumed to be normally and independently distributed with variance $\sigma_i^2 = h(z_i'\alpha)$ where $h$ denotes a function, and $z_i'$ is a $1 \times P$ vector of variables influencing heteroskedasticity.

Let $Z$ be an $N \times P$ matrix with row vectors $z_i'$. Some of the variables in $Z$ could be the same as the variables in $X$.

Regress $e^2/\sigma_{ML}^2$ on $Z$, including an intercept term.

Note that (sum of squares due to $Z)/2 \sim \chi^2(P-1)$ approximately. The factor $1/2$ appears here since under normality the variance of $\varepsilon^2/\sigma^2$ is $2(E\varepsilon^4 = 3\sigma^4)$.

# Testing for heteroskedasticity (cont'd):

An alternative approach (Koenker) drops normality and estimates the variance of $e_i^2$ directly by $N^{-1} \sum (e_i^2 - \hat{\sigma}^2)^2$. The resulting statistic can be obtained by regressing $e^2$ on $z$ and looking at $NR^2$ from this regression.

Other tests are available for time series.

# Testing Normality

The moment generating function of a random variable $x$ is $m(t) = E(\exp(tx))$; note $m'(0) = Ex$; $m''(0) = Ex^2$; etc.
The MGF of the normal distribution $n(\mu, \sigma^2)$ is $m(t) = \exp(t\mu + t^2\sigma^2/2)$.

*Proof*:
let $c = (2\pi\sigma)^{-1/2}$

$$
\begin{aligned}
m(t) &= c\int \exp(tx)\exp(-1/2(x-\mu)^2/\sigma^2)dx \\
&= c\int \exp(-1/2(x-\mu-\sigma^2 t)^2/\sigma^2 + t\mu + \sigma^2 t^2/2)dx \\
&= \exp(t\mu + \sigma^2 t^2/2).
\end{aligned}
$$

# Testing Normality (cont'd)

Thus for the regression errors $\varepsilon$ we have
$E\varepsilon = 0$; $E\varepsilon^2 = \sigma^2$; $E\varepsilon^3 = 0$; $E\varepsilon^4 = 3\sigma^4$; $E\varepsilon^5 = 0$; etc.

It is easier to test the 3rd and 4th moment conditions than normality directly.
If we knew the $\varepsilon$, it would be easy to come up with a $\chi^2$ test.

In fact a test can be formed using the residuals $e$ instead (and relying on asymptotic distibution theory). The test statistic is

$$n[\overline{((e/s)^3)^2}/6 + (\overline{(e/s)^4} - 3)^2/24].$$

Which is $\chi^2$ with 2 $df$.

This is the Kiefer/Salmon test (also called Jarque/Bera).