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 β_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 β_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 σ^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 $X^* = [X_1, X_2]$, where X_1 is $N \times K_1$ and X_2 is $N \times K_2$. That is, the "true" model is

$$Y = X_1\beta_1 + \varepsilon.$$

Proposition: $b^* = [b_1^{*\prime}, b_2^{*\prime}]'$ where

$$b_1^* = (X_1' M_2 X_1)^{-1} X_1' M_2 Y,$$
$$b_2^* = (X_2' M_1 X_2)^{-1} X_2' M_1 Y.$$

Application 2: Inclusion of unnecessary variables (cont'd)

Proof: Premultiplying both sides of the estimated model

$$Y = X_1 b_1^* + X_2 b_2^* + e^*$$

by M_2 we get:

$$M_2 Y = M_2 X_1 b_1^* + e^*$$

because $M_2 e^* = e^*$ (why?). Premultiplying both sides of the above equation by X_1' we get:

$$X_1' M_2 Y = X_1' M_2 X_1 b_1^* \Longrightarrow b_1^* = (X_1' M_2 X_1)^{-1} X_1' M_2 Y$$

The same reasoning applies to b_2^* as well.

Comment: the above is the same as regressing X_1 on X_2 and then regressing Y on the residuals of the former regression. Very neat interpretation!

Application 2: Inclusion of unnecessary variables (cont'd)

Proposition: b^* is unbiased.
Proof: Substituting $Y = X_1\beta_1 + \varepsilon$ in b_1^* and b_2^* we get:

$$b_1^* = \beta_1 + (X_1' M_2 X_1)^{-1} X_1' M_2 \varepsilon$$
$$b_2^* = (X_2' M_1 X_2)^{-1} X_2' M_1 \varepsilon$$

from which it easily follows that $\mathrm{E}(b_1^*) = \beta_1$ and $\mathrm{E}(b_2^*) = 0$.

Proposition: $\mathrm{V}(b_1^*) \geq \mathrm{V}(b_1)$, where $b_1 = (X_1' X_1)^{-1} X_1' Y$.
Proof: From the equation of b_1^* we get
$\mathrm{V}(b_1^*) = \sigma^2 (X_1' M_2 X_1)^{-1} \Longrightarrow \mathrm{V}(b_1^*) \geq \mathrm{V}(b_1)$ because

$$\sigma^{-2}(\mathrm{V}(b_1)^{-1} - \mathrm{V}(b_1^*)^{-1}) = X_1' X_1 - X_1' M_2 X_1 = X_1' P_2 X_1 = X_1' P_2' P_2 X_1$$

where $P_2 = X_2 (X_2' X_2)^{-1} X_2'$.

Estimation of Variance

Under normality of the errors ε, since $e^* = M^* Y = M^* \varepsilon$, where $M^* = I_N - X^*(X^{*\prime\prime} X^*)^{-1} X^{*\prime}$, we get

$$e^{*\prime} e^* = \varepsilon M^* \varepsilon = \sigma^2 \chi^2 (N - K_1 - K_2)$$

whereas

$$e^{\prime} e = \varepsilon M_1 \varepsilon = \sigma^2 \chi^2 (N - K_1)$$

from which it follows that

$$\mathrm{E}\left(\frac{e^{*\prime} e^*}{N - K_1 - K_2} \right) = \mathrm{E}\left(\frac{e^{\prime} e}{N - K_1} \right) = \sigma^2 .$$

Heteroskedasticity

Suppose $V = \mathrm{V}(\varepsilon) = \{v_{ij}\}$ with $v_{ij} = 0$ for any $i \neq j$.
Is the OLS estimator unbiased? Is it BLUE?

Proposition: Under the assumption of heteroskedasticity,

$$\mathrm{V}(\widehat{\beta}) = (X'X)^{-1}X'VX(X'X)^{-1}$$

Proof:

$$\mathrm{V}(\widehat{\beta}) = \mathrm{E}(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1} = (X'X)^{-1}X'VX(X'X)^{-1}\blacksquare$$

Note that

$$X'VX = \mathrm{E}\sum_{i=1}^{N} X_i\varepsilon_i^2X_i'$$

where $X' = [X_1, ..., X_N]$. The above suggests to estimate $X'VX$ with

$$\sum_{i=1}^{N} X_i e_i^2 X_i'$$

when V is diagonal.

Testing for heteroskedasticity:

1. Goldfeld-Quandt test:

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

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

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 ε_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/σ_{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 ε^2/σ^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 ε 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 ε, it would be easy to come up with a χ^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 χ^2 with 2 df.

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