void

Multicollinearity

True model

$$y = b_0 + b_1 X_1 + \dots + b_k X_k + u_t$$

1. Perfect collinearity between X_k and X_1 means we can only estimate sum of the effects; no separate identification. i.e.

$$X_k = a X_1 \Rightarrow$$

$$y = b_0 + (b_1 + a b_k) X_1 + b_2 X_2 + \dots + b_{k-1} X_{k-1} + u_t$$

so we identify $\hat{b} = b_1 + a b_k$ but not b_1 and b_k.

2. Imperfect collinearity raises the est. standard errors of the estimators. i.e.

$$\sigma^2_{\hat{b}_k} = \frac{\sigma^2_u}{\sum v^2_{kt}}$$

where v_{kt} is independent variation in X_{kt}. $\sigma^2_{\hat{b}_k} \uparrow$ as $v_{kt} \downarrow$.

Symptom: Reasonable explanatory power of the regression together w/ generally insignificant t-ratios on individual variables.

Imp: The problem of collinearity is insufficient information in the sample. The solution is more information. This can come from two "sources"

1. More data.

$\sigma_b^2 \downarrow$ as $T \uparrow$ because $\sum_1^T \hat{v}_t^2 \uparrow$.

2. A priori information

Eg. of a priori info

Consider C.D. prod. fn.

$$Q = A L^{b_1} K^{b_2} e^{U}$$

$$q = a + b_1 l + b_2 k + U_t$$

Suppose l and k are highly collinear. If we assume CRS, $b_2 = 1 - b_2$ and we estimate the restricted form:

$$q - l = a + b_2 (k - l) + U_t$$

Then we retrieve $\hat{b}_1 = 1 - \hat{b}_2$. This "solves" the collinearity problem by adding info on b_2 which we assume, but do not test, i.e. $b_1 = 1 - b_2$. Of course, if the p.f. does not exhibit CRS we have a misspecified model and biased coefficients.

Remember: The Gauss-Markov theorem holds regardless of multicollinearity. Hence, estimators are still BLUE. Predictions based on those estimators are also BLU predictions.

Notes on Economic Interpretations of Regressions: Elasticities

i) Let

$$y = a_0 + a_1 x_1 + a_2 x_2 + U$$

Define elasticity of y wrt x_i:

$$\varepsilon_i = \frac{\partial y}{\partial x_i} \frac{x_i}{y}$$

Then

$$\varepsilon_i = a_i \frac{x_i}{y}$$

which varies with the level of x_i and y. So we must evaluate elasticity at a specified level of x_i, y.

ii) Now consider log linear formulation:

$$\log y = \alpha_0 + \alpha_1 \log x_1 + \alpha_2 \log x_2 + v$$

$$\varepsilon_i = \frac{\partial y}{\partial x_i} \frac{x_i}{y_i}$$

$$\alpha_i = \frac{\partial \log y}{\partial \log x_i} \equiv \varepsilon_i$$

Hence, log linear formulation implies a priori constant elasticities.

Lecture on Heteroscedasticity

We assumed so far that error were homoscedastic:

$$E(U_t^2) = \sigma_U^2 \quad \forall t.$$

Now we let $E(U_t^2)$ depend on some variables. As an example consider the consumption function seen before:

$$C_t = b_0 + b_1 y_{dt} + b_2 A_{t-1} + U_t$$

with $E(U_t^2) = y_{dt}\sigma_U^2$, $E(U_t U_s) = 0 \quad t \neq s$.
We maintain the assumption $E(U_t \mid y_{dt}, A_{t-1}) = 0$.
This means:

$$E(U_t y_{dt}) = E(U_t A_{t-1}) = 0$$

but: $$E(U_t^2) = y_{dt}\sigma_U^2$$

Consequences for Estimators

1. We showed

$$\hat{b} = \sum w_t u_t + b$$

so: $E(\hat{b}) = b$ unbiasedness is preserved

as: $= \sum w_t E(u_t) + b$

2. $E(\hat{b}-b)^2 = \sum w_t^2 E(u_t^2) \neq \sigma_u^2 \sum w_t^2$

so variance formulas are invalid

Estimation Procedure : GLS

Consider dividing thru the model by $\sqrt{y_{dt}}$:

$$\frac{c_t}{\sqrt{y_{dt}}} = b_0\left(\frac{1}{\sqrt{y_{dt}}}\right) + b_1 \frac{y_t}{\sqrt{y_{dt}}} + b_2 \frac{A_{t-1}}{\sqrt{y_{dt}}} + \frac{u_t}{\sqrt{y_{dt}}}$$

$$c_t^* = b_0^* + b_1 y_{dt}^* + b_2 A_{t-1}^* + u_t^*$$

Note:

$$E(U_t^*) = \frac{1}{\sqrt{y_{dt}}} E(U_t) = 0$$

$$E\left(U_t^* \frac{1}{\sqrt{y_{dt}}}\right) = E(U_t^* \sqrt{y_{dt}}) = E(U_t^* A_{t-1}/\sqrt{y_{dt}}) = 0$$

$$E(U_t^{*2}) = \frac{1}{y_{dt}} E(U_t^2) = \frac{y_t \sigma_u^2}{y_{dt}} = \sigma_u^2$$

Hence, the transformed model in c^*, y^*, A^* has disturbances which satisfy orthogonality, zero mean and constant variance. Hence, we can apply OLS to the transformed model.

All standard formulas apply, using starred variates in place of original ones.

Intuition: Each observation is subject to random noise. If some observations are subject to more noise than others, we should give them "more weight" in the regression because they contain "more information." The weight we give

to each observation is inversely proportional to $\sqrt{y_{dt}}$, where $E(u_t^2) = y_{dt}\sigma_u^2$.

Extensions

1. Convert, or express, the model in a form where homoskedastic errors are more plausible.

a) For example, instead of a model relating profits to asset size, -

(i) $\pi_t = b_0 + b_1 A_t + u_t$

where we expect $E(u_t^2) = f(A_t)\sigma_u^2$, consider relating profit rates to size,

(ii) $\pi_t^* = b_0' + b_1 A_t + u_t'$

where $\pi_t^* = \pi_t / A_t$. Note, though, that the two models above do differ, because if (ii) is fine, then absolute profits depend on A_t and A_t^2 with no constant, and (i) is misspecified.

b) Another example, logarithmic transforms often help to eliminate heteroskedastic errors because relation is in proportional Terms.

$$Q_t = a_0 + a_1 L_t + a_2 K_t + U_t$$

$$q_t = b_0 + b_1 l_t + b_2 k_t + u_t$$

We expect $E(u_t^2) = \sigma_u^2$ more likely than $E(U_t^2) = \sigma_U^2$ because log transformation reduces variation in length (shrinks the scale).

2. Polynomial Approximation To Heteroskedastic errors

Suppose we are willing to assume:

$$E(u_t^2) = \sigma_u^2 f(X_{2t})$$

in $y_t = b_0 + b_1 X_{1t} + b_2 X_{2t} + U_t$

We can think of decomposing U_t^2 into a part we can express as a function of X_{2t} and an unobserved

part:

$$u_t^2 = f(x_{2t}) + \varepsilon_t \quad \text{with} \quad E(\varepsilon_t x_{2t}^k) = 0 \quad \forall k.$$

Expand $f(x_{2t})$ to k powers

$$u_t^2 = a_0 + a_1 x_{2t} + a_2 x_{2t}^2 + \dots + a_k x_{2t}^k + \varepsilon_t$$

where $E(\varepsilon_t x_{2t}^k) = 0 \ \forall k$ shows this to be linear regression equation. Then if we estimate this eq, obtain

$$\widehat{f(x_{2t})} = \hat{a}_0 + \hat{a}_1 x_{2t} + \dots + \hat{a}_k x_{2t}^k$$

and define

$$y_t^* = y_t \Big/ \sqrt{\widehat{f(x_{2t})}}$$

$$x_{1t}^* = x_{1t} \Big/ \sqrt{\widehat{f(x_{2t})}}$$

$$x_{2t}^* = x_{2t} \Big/ \sqrt{\widehat{f(x_{2t})}}$$

$$u_t^* = u_t \Big/ \sqrt{\widehat{f(x_{2t})}}$$

Then estimate by OLS

$$y_t^* = b_0 \frac{1}{\sqrt{\hat{f}}} + b_1 x_{1t}^* + b_2 x_{2t}^* + u_t^*$$

To test for heteroskedasticity, estimate

$$\hat{u}_t^2 = a_0 + a_1 x_{2t} + \cdots + a_k x_{kt}^k + \varepsilon_t$$

and run an F-test on the hypothesis $a_1 = a_2 = \cdots = a_k =$