Problems with Errors

Quantitative Methods II for Political Science
Kenneth Benoit

February 11, 2009
Non-zero expected error problems

- **Constant non-zero mean**
  - Happens when there are systematically positive or negative errors of measurement in the dependent variable
  - Consequence: the OLS estimate of the intercept will be biased

- **Zero intercept**
  - No bias from including an unnecessary intercept, even when theory suggests it should be zero
  - Not including an intercept is equivalent to the linear constraint that $\beta_0 = 0$

- **Non-constant error variances**, aka heteroskedasticity

- **Limited dependent variable**
  - In this special case, OLS will be biased for all coefficients
  - We will deal with this more in Week 6, since it requires non-OLS solutions
“Non-spherical” error

- Means that the variance of the residuals is not uniform, OR
- Means that the residuals may be correlated

\[ V_c \text{ is } n \times n = \begin{bmatrix} v(e_1) & cov(e_1,e_2) & v(e_2) \\ cov(e_1,e_2) & v(e_2) & v(e_3) \\ \vdots & \vdots & \ddots \\ v(e_{n-1}) & v(e_{n-1}) & v(e_n) \end{bmatrix} \]

- Consequences
  - Efficiency loss
  - Inconsistency: can no longer trust $\beta_{OLS}$
  - $\beta_{OLS}$ is no longer the maximum likelihood estimator
Heteroskedasticity

- Graphical inspection of residuals is best check
- Some tests also exist: Breusch-Pagan test, Goldfeld-Quandt test (see the \texttt{lmtest} library, it contains all of these)
- One solution is to use generalized least squares (GLS)
- A fix: White's heteroskedasticity-corrected standard errors:

\[ \text{Var}(b) = (X'X)^{-1}X'diag(e^2)X(X'X)^{-1} \]

- Can implement these using \texttt{hccm()} from the \texttt{car} package
Autocorrelated disturbances

- **Spatial autocorrelation**: Caused when a shock in one period affects shocks in a subsequent period

- In time-series data, shocks often have effects that persist for more than one period
- Can test for this using the Durbin-Watson test, which tests the first order autocorrelation coefficient $\rho$
  - $0 < d < 4$
  - $d = 2$ when $\rho = 0$; if $d < 1.0$ then need correction
  - use `dwtest()` from the `testlm` package
Simultaneous Equations: The Problem

Assume we have the following model:

- \( E = \) evaluations of parties
- \( P = \) party identification
- \( V = \) vote to be explained
- \( E = q_1 P + U_1 \) (1)
- \( V = q_2 P + q_3 E + U_2 \) (2)

Question: What is the total effect of party ID on voting behaviour? Includes:

- the direct effect \( q_2 \), plus
- the indirect effect \( q_1 q_3 \)

Substitute (1) into (2):

\[
V = q_2 P + q_3 (q_1 P + U_1) + U_2 \\
= q_2 P + q_3 q_1 P + q_3 U_1 + U_2 \\
= (q_2 + q_3 q_1) P + q_3 U_1 + U_2 \\
= \pi_i P + \nu
\]
Simultaneous Equations: The Problem

▶ Path model:

“Path analysis”: uses correlations instead of covariances, so that all estimated relationships are standardized coefficients.
General Multiequation Model

\[
Y_1 = \beta_{11} X_1 + \beta_{21} X_2 + \beta_{31} X_3 + \beta_{41} X_4 + U_1
\]

\[
Y_2 = \gamma_{12} Y_1 + \beta_{12} X_1 + \beta_{22} X_2 + \beta_{32} X_3 + \beta_{42} X_4 + U_2
\]

\[
Y_3 = \gamma_{13} Y_1 + \gamma_{23} Y_2 + \beta_{13} X_1 + \beta_{23} X_2 + \beta_{33} X_3 + \beta_{43} X_4 + U_3
\]

- where \( Y_m \) are \textit{endogenous} variables, \( X_k \) are \textit{exogenous} variables
- Hierarchical equations are structured so that higher ordered endogenous variables do not appear as explanatory variables in lower ordered equations
- Structural equations since they represent underlying systematic and stochastic processes which led to the observed data
- To access the \textit{total effect} on the endogenous variables of a change in \( X \), we must include all indirect effects through \( \gamma \) and \( \beta \)
- We do this through the \textit{reduced form equations}
Reduced form equations 1

\[ Y_1 = \beta_{11} X_1 + \beta_{21} X_2 + \beta_{31} X_3 - \beta_{41} X_4 + u_1 \]

\[ Y_1 = \pi_{11} X_1 + \pi_{21} X_2 + \pi_{31} X_3 + \pi_{41} X_4 + v_1 \]
Reduced form equations 2

\[ Y_2 = \gamma_{12} (\beta_{11} X_1 + \beta_{21} X_2 + \beta_{31} X_3 + \beta_{41} X_4 \theta + \varphi) + \alpha_1 + \beta_{12} X_1 + \beta_{22} X_2 + \beta_{32} X_3 + \beta_{42} X_4 \gamma + \varphi \times X_4 + \epsilon_2 \]

\[ = (\beta_{12} + \gamma_{12} \beta_{11}) X_1 + (\beta_{22} + \gamma_{12} \beta_{21}) X_2 + (\beta_{32} + \gamma_{12} \beta_{31}) X_3 + (\beta_{42} + \gamma_{12} \beta_{41}) X_4 \gamma + \epsilon_2 + \epsilon_2 \varphi \]

\[ Y_2 = \Pi_{12} X_1 + \Pi_{22} X_2 + \Pi_{32} X_3 + \Pi_{42} X_4 \gamma + \epsilon_2 \]
\[
Y_3 = \left[ \beta_{13} + \gamma_{23} \beta_{12} + (\gamma_{23} \gamma_{12} + \gamma_{13}) \beta_{11} \right] X_1 + \\
\left[ \beta_{23} + \gamma_{23} \beta_{22} + (\gamma_{23} \gamma_{12} + \gamma_{13}) \beta_{21} \right] X_2 + \\
\left[ \beta_{33} + \gamma_{23} \beta_{32} + (\gamma_{23} \gamma_{12} + \gamma_{13}) \beta_{31} \right] X_3 + \\
\left[ \beta_{43} + \gamma_{23} \beta_{42} + (\gamma_{23} \gamma_{12} + \gamma_{13}) \beta_{41} \right] X_4 + \\
\left[ \beta_{53} + \gamma_{23} \beta_{52} + (\gamma_{23} \gamma_{12} + \gamma_{13}) \beta_{51} \right] X_5
\]
Using reduced form equations to gauge total effect

Question: What is the total effect on $Y_3$ of a change in $X_2$?

\[
\begin{align*}
\beta_{23} & \quad \text{direct effect} \\
\gamma_{23}\beta_{22} & \quad \text{indirect effect from direct change in } Y_2 \\
\gamma_{13}\beta_{21} & \quad \text{indirect effect from a direct change in } Y_1 \\
+ \quad \gamma_{23}\gamma_{12}\beta_{21} & \quad \text{indirect effect due to changes in } Y_2 \text{ caused by changes in } Y_1 \\
\pi_{23} & \quad \text{total effect}
\end{align*}
\]
Why are structural equations needed?

- If exogenous variables are independent of error terms, then we can use OLS to estimate unbiased and consistent estimates of reduced form parameters. But estimates of the structural parameters will be biased and inconsistent. So?

- Generally in political science we are concerned with underlying causal relationships — in other words, the structural parameters — to test competing theories. Example: electoral model

- Multiequation models that are hierarchical and have independent error terms across equations are called recursive systems
  - for recursive systems, we can use OLS
  - for non-recursive systems, OLS causes bias and inconsistency

- Key question: are the error terms independent?
The independence of error terms

- This is the key question
- Is the uncorrelated error term assumption reasonable?
  - Errors may be the result of omitted small influences that could be similar across equations
  - If some explanatory factor is excluded from more than one equation, this will cause correlated errors
  - If some $X$ has measurement error, this can also cause errors that correlate across equations
- The Hausman test (or Durbin-Wu-Hausman test—see Kennedy) can be used to test the assumption of endogeneity
  - regress the the endogenous variable on the instruments (typically, all exogenous variables)
  - then regress the main dependent variable on the exogenous variables plus the residuals from step 1
  - the $t$-test on the residual coefficient is the test for endogeneity
Consider the following hierarchical model, where the error terms are correlated:

\[
Y_1 = \beta_{11}X_1 + \beta_{21}X_2 + \beta_{31}X_3 + U_1, \quad (8.19)
\]

\[
Y_2 = \gamma_{12} Y_1 + \beta_{12}X_1 + \beta_{32}X_3 + \beta_{42}X_4 + U_2, \quad (8.20)
\]

\[
Y_3 = \gamma_{13} Y_1 + \gamma_{23} Y_2 + \beta_{13}X_1 + U_3, \quad (8.21)
\]
The reduced form expressions for each endogenous variable are:

\[ Y_1 = \beta_{11}X_1 + \beta_{21}X_2 + \beta_{31}X_3 + U_1 = \pi_{11}X_1 + \pi_{21}X_2 + \pi_{31}X_3 + V_1, \quad (8.22) \]

\[ Y_2 = (\beta_{12} + \gamma_{12} \beta_{11})X_1 + \gamma_{12} \beta_{21}X_2 + (\beta_{32} + \gamma_{12} \beta_{31})X_3 + \beta_{42}X_4 + \gamma_{12}U_1 + U_2 = \pi_{12}X_1 + \pi_{22}X_2 + \pi_{32}X_3 + \pi_{42}X_4 + V_2, \quad (8.23) \]

\[ Y_3 = (\beta_{13} + \gamma_{13} \beta_{11} + \gamma_{23} \beta_{12} + \gamma_{23} \gamma_{12} \beta_{11})X_1 + (\gamma_{13} \beta_{21} + \gamma_{23} \gamma_{12} \beta_{21})X_2 + (\gamma_{13} \beta_{31} + \gamma_{23} \beta_{32} + \gamma_{23} \gamma_{12} \beta_{31})X_3 + \gamma_{23} \beta_{42}X_4 + (\gamma_{13} + \gamma_{23} \gamma_{12})U_1 + \gamma_{23}U_2 + U_3 = \pi_{13}X_1 + \pi_{23}X_2 + \pi_{33}X_3 + \pi_{43}X_4 + V_3. \quad (8.24) \]
Solution 1: Indirect Least Squares

- We directly estimate $\beta_{11}$, $\beta_{21}$, $\beta_{31}$ from the regression on $Y_1$(8.22)
- This leaves 3 “unknowns”: $\gamma_{12}$, $\beta_{12}$, $\beta_{32}$

\[
\begin{align*}
\pi_{22} &= \gamma_{12}\beta_{21} \\
\pi_{12} &= \beta_{12} + \gamma_{12}\beta_{11} \\
\pi_{32} &= \beta_{32} + \gamma_{12}\beta_{31}
\end{align*}
\]

- We can estimate $\gamma_{12}$ using the previous estimate for $\beta_{21}$
- and then we can use $\beta_{11}$, $\gamma_{12}$ to estimate $\beta_{12}$
- and use $\beta_{31}$, $\gamma_{12}$ to estimate $\beta_{31}$
- This method is known as **indirect least squares**
Problem: indirect least squares does not work on 8.24

\[ \gamma_{23} = \frac{\pi_{43}}{\beta_{42}} = \frac{\pi_{43}}{\pi_{42}} \quad \text{OK} \]

but:

\[ \gamma_{13} = \frac{(\pi_{23} - \gamma_{23}\pi_{22})}{\pi_{21}} \]
\[ \gamma_{13}^* = \frac{(\pi_{33} - \gamma_{23}\pi_{32})}{\pi_{31}} \]

in finite samples, \( \gamma_{13} \neq \gamma_{13}^* \)

This is known as “overidentification” and comes from having 4 expressions for 3 unknowns
Find an appropriate instrumental variable for each endogenous variable — these are known as “instrumental variables”

The IVs will act as substitutes for explanatory variables that are correlated with the explanatory variable, but uncorrelated with the error term

Obtain:

\[ \hat{Z}_2 = \hat{Y}_1 \] = reduced form

\[ \hat{Z}_3 = \hat{Y}_2 \] = reduced form

Regress \( Y_3 \) on \( X_1, \hat{Z}_2, \) and \( \hat{Z}_3 \) to estimate 8.21

Not unbiased, but it is consistent
Solution 3: Two-stage least squares

- A special case of the IV approach: combines all exogenous variables to create a “best” IV
- Regress each endogenous variable (which are on RHS) on all of the exogenous variables in the system, and use the estimated values of each endogenous variable from these regressions as IVs
  1. Regress each endogenous variable (that is a regressor) on all exogenous variables in the system of simultaneous equations, and calculated the estimated values of the endogenous variables
  2. Use the estimated values of the endogenous variables as IVs
- This can be done “manually” in steps, or using the `tsls()` command in R (requires the `sem` library)