Problems with Errors

Linear Regression Analysis
Kenneth Benoit

August 20, 2012

Regression diagnostics in Stata

- Following `regress`, there are regress postestimation quantities that can be generated directly using `predict`:

`, xb` generates $\hat{y}_i$

`, residuals` generates e_i

`, rstudent` generates studentized residual for i

`, cooksd` generates Cook's distance for i

`, leverage` generates leverage of obsveration i

- And several diagnostic plots

`rvfplot`

`plot ¡varname¿`

`lvr2plot`

Studentized and standardized residuals

- Define h_i as the diagonal H_{ii} from the hat matrix. These are also known as the leverage of each observation, and is computed as:

$$h_i = x_i(X'X)^{-1}x_j'$$

- The standardized residual is then:

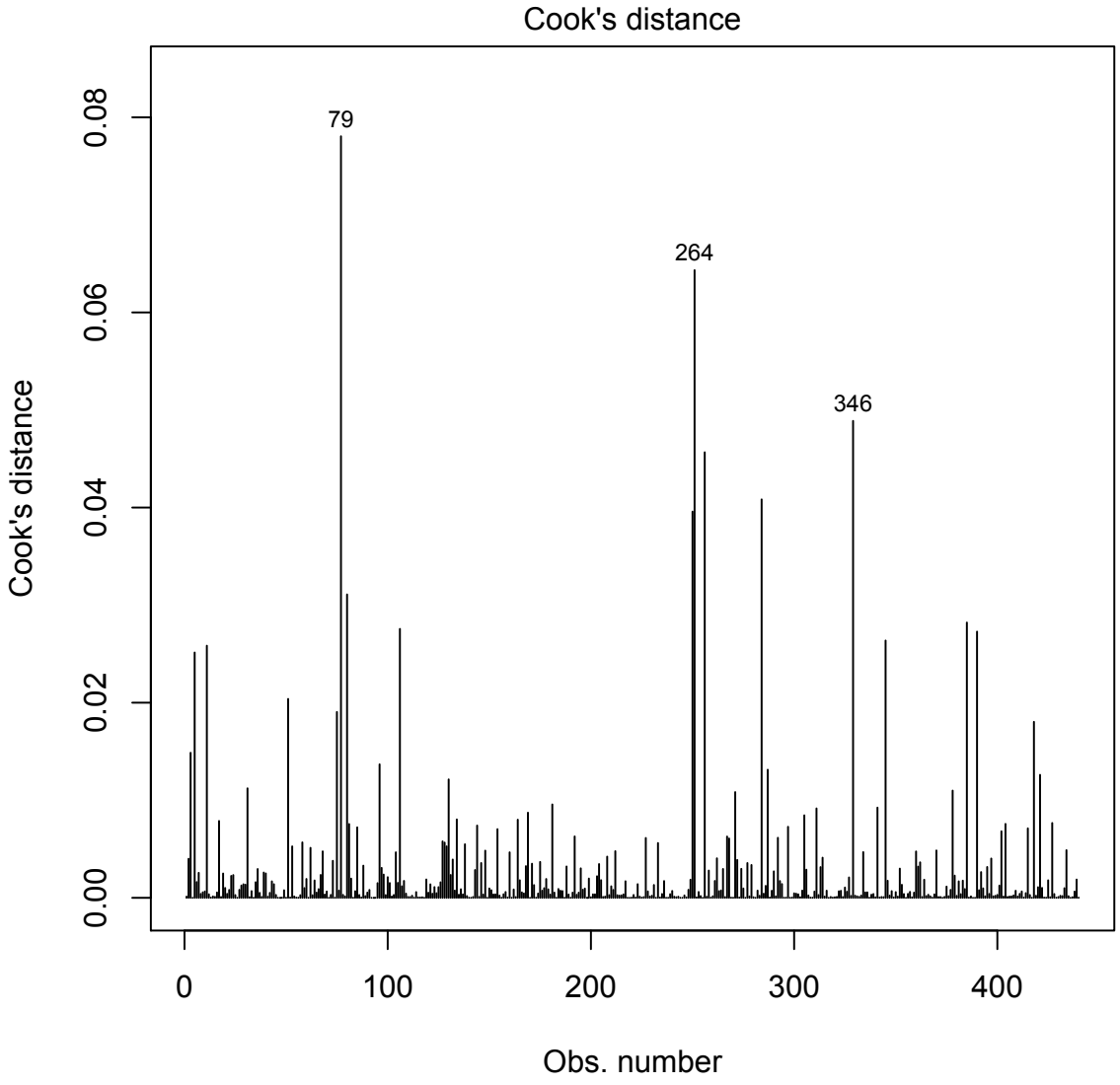
$$\hat{e}_i = \frac{e_i}{s\sqrt{1-h_i}}$$

- The studentized residual is the root mean squared error of the refression with the ith observation removed:

$$r_i = \frac{e_i}{s_{(i)}\sqrt{1-h_i}}$$

- Both standardized and studentized residuals are attempts to adjust residuals by their standard errors, where the $\text{Var}(e_i) = \sigma^2(1-h_i)$
- Note that the calculated $e_i = Y_i - \hat{Y}_i$ all have the same variance (the homoskedasticity assumption), but the calculated e_i do not

Cook's Distance plot



- Cook's Distance is a measure of *influence*

```
plot(lm(votes1st~spend_total*incumb, data=dail), which=4)
```

Cook's Distance

- Cook's distance for observation i measures the effect of deleting that observation
- Defined as:

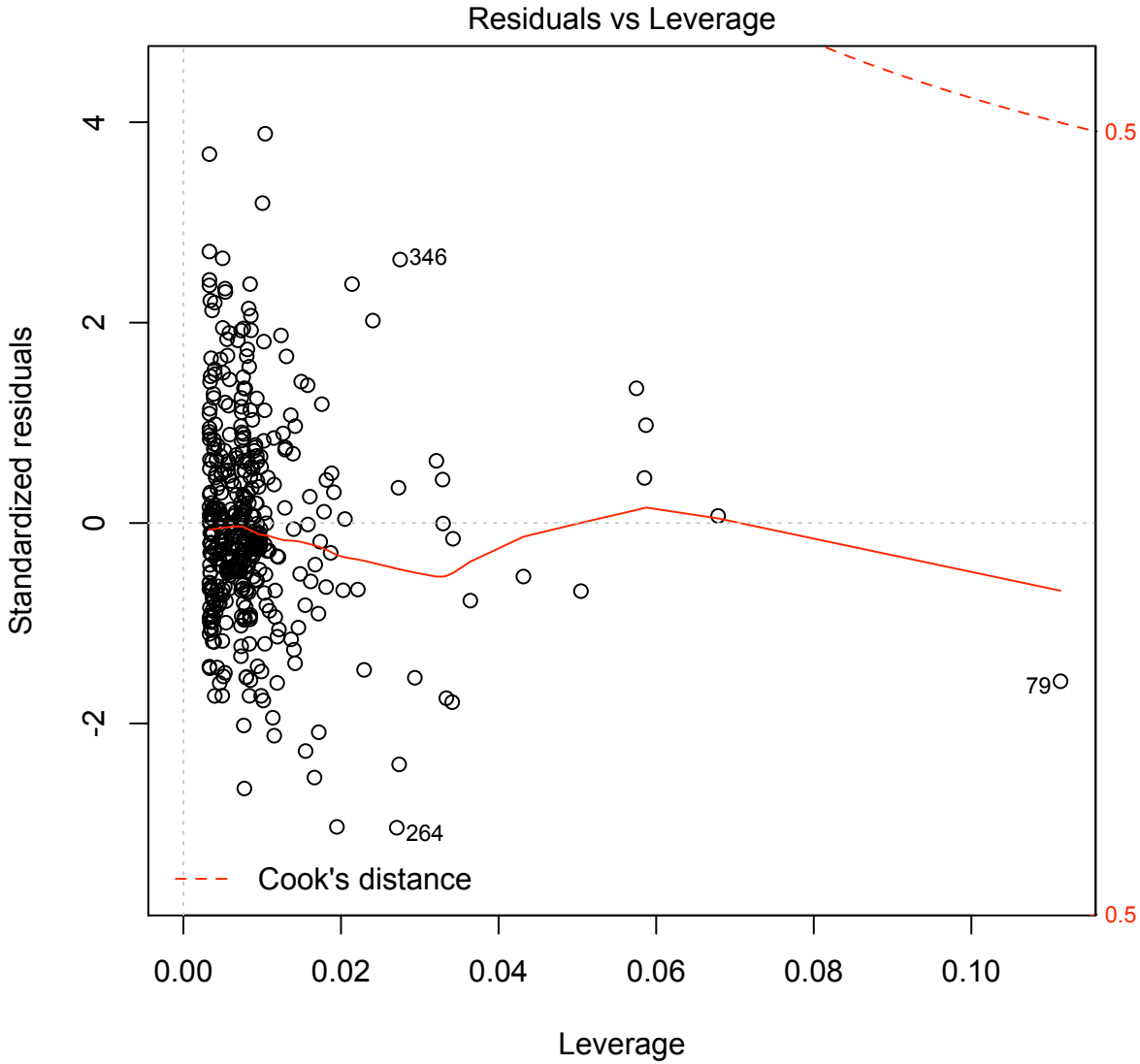
$$\begin{aligned} D_i &= \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{p\hat{\sigma}} \\ &= \frac{e_i^2}{p\hat{\sigma}}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right] \end{aligned}$$

- Guideline: points for which $D_i > 1.0$ usually need closer examination
- Closely connected to *leverage*, which is h_{ii} or the diagonal of the hat matrix H

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{s_{xx}}$$

- In R, `hii <- influence()$hat`

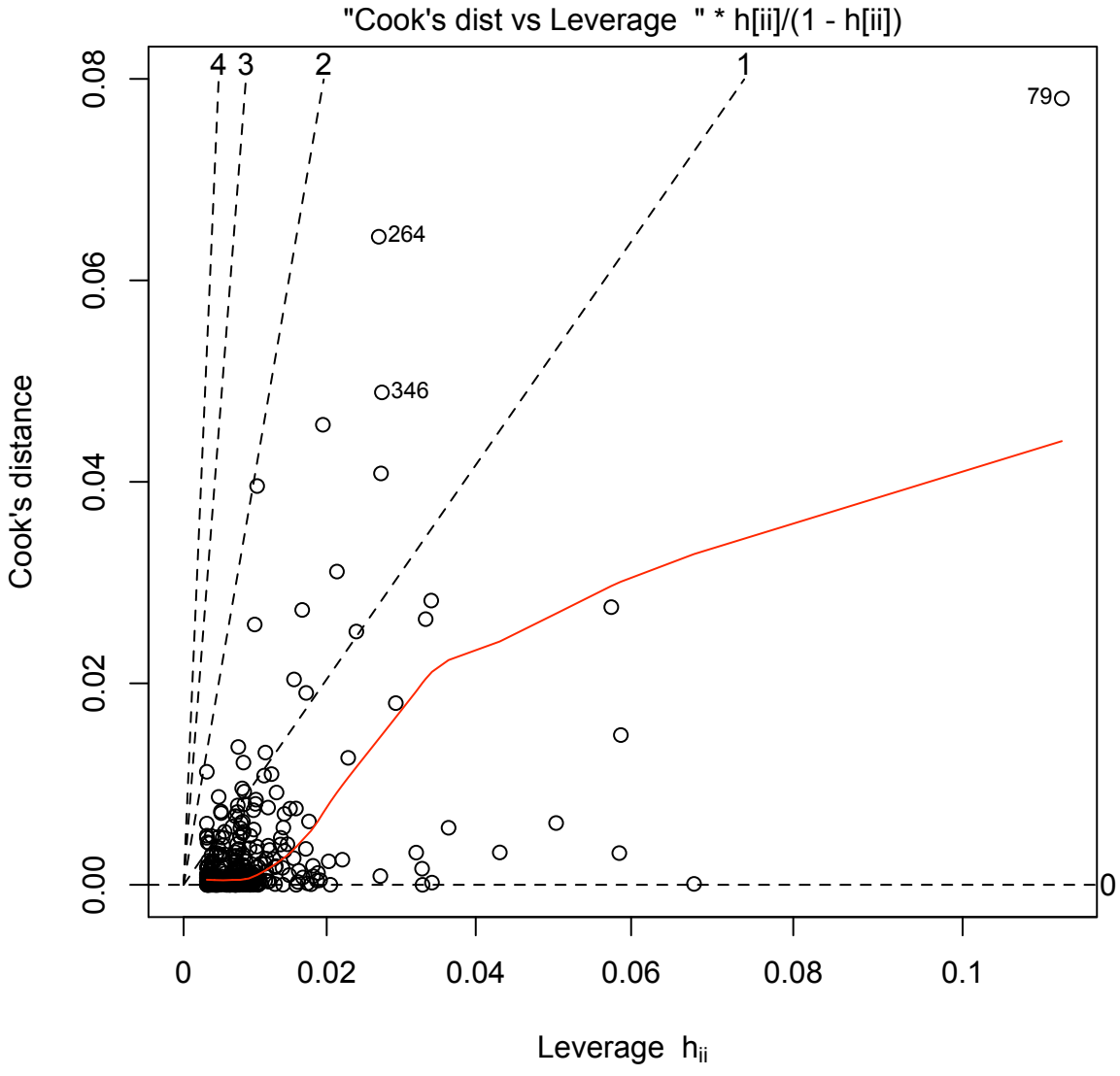
Residuals v. leverage



- Leverage is defined as h_{ii} for the ith observation

```
plot(lm(votes1st~spend_total*incumb, data=dail), which=5)
```

Cook's Distance v. leverage




```
plot(lm(votes1st~spend_total*incumb, data=dail), which=6)
```

What to do with Outliers?

1. Ignore the problem!
2. Investigate *why* the data are outliers — what makes them unusual?
3. Consider respecifying the model, either by tranforming a variable or by including an additional variable (but beware of overfitting)
4. Consider a variant of "robust regression" that downweights outliers
5. Note: Fox contains a number of good rules of thumb for detecting what constitutes an "outlier" based on thresholds applied to diagnostic statistics (but best method is usually graphical)

Robust regression methods

- Robust regression methods are methods to find central tendencies without giving undue influence to outliers
- For simple case like $\bar{X}$, assume we have extreme values
- Simple solution: Trim off extreme values, such as the outer deciles: $\bar{X}_{.10}$ is the mean of X value with outer 10% trimmed away
- Or: use $\bar{X}_b$ as "biweighted" mean, gradually weighting observations

$$\begin{aligned} w &= (1-z^2)^2 \text{ if } |z| \leq 1 \\ &= 0 \text{ if } |z| > 1 \\ \text{where } z &= \frac{X - \text{median}(X)}{3(\text{IQR})} \\ \text{and } \bar{X}_b &= \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i} \end{aligned}$$

- The "bi-weighted iterative" mean takes this a step further:

$$z = \frac{X - \bar{X}_b}{3(\text{IQR})} \tag{1}$$

Biweighted least squares

$$
\begin{aligned}
e_i &= Y_i - \hat{Y}_i \\
z_i &= \frac{Y_i - \hat{Y}_i}{3s} \\
w &= (1 - z^2)^2 \text{ if } |z| \leq 1 \\
&= 0 \text{ if } |z| > 1 \\
b &= \frac{\sum wXY}{\sum wX^2} \\
a &= \bar{Y} - b\bar{X}
\end{aligned}
$$

1. Start with OLS (same as all $w_i = 1$)
2. Measure $Y_i - \hat{Y}_i$ — this determines the weights w_i
3. Fit the new line and repeat steps 2–3 until improvement stops

   In R, this is `rlm()` from the `MASS` library

Changes of scale

- In short: Linear rescaling of variables will not change the essential key statistics for inference, just their scale
- Suppose we reexpress x_i as $(x_i + a)/b$. Then:
  - t, F, $\hat{\sigma}^2$, R^2 unchanged
  - $\hat{\beta}_i \rightarrow b\hat{\beta}_i$
- Suppose we rescale y_i as $(y_i + a)/b$. Then:
  - t, F, R^2 unchanged
  - $\hat{\sigma}^2$ and $\hat{\beta}_i$ will be rescaled by b
- Standardized variables and standardized coefficients: where we replace the variables (all x and y) by their standardized values $(x_i - \bar{X})/\text{SD}_x$ (e.g. for x). Standardized coefficients are sometimes called "betas".

More on standardized coefficients

Consider a standardized coefficient b^* on a single variable x.

- Formula: $b^* = b\frac{\text{SD}_x}{\text{SD}_y}$
- Intrepretation: the increase in standard deviations of y associated with a one standard deviation increase in x
- where standardization means transforming into variables z such that $z_i = (x_i - \bar{X})/\text{SD}_x$
- Motivation: "standardizes" units so we can compare the magnitude of different variables' effects
- In practice: serious people never use these and you should not either
  - too tricky to interpret
  - misleading since suggests we can compare apples and oranges
  - too dependent on sample variation (just another version of R^2)

Collinearity

- When some variables are exact linear combinations of others then we have exact collinearity, and there is no unque least squares estimate of β
- When X variables are correlated, then we have (multi)collinearity
- Detecting (multi)collinearity:
  - look at correlation matrix of predictors for *pairwise* correlations
  - regress x_k on all other predictors to produce R_k^2, and look for high values (close to 1.0)
  - Examine eigenvalues of $X'X$

Collinearity continued

- Define:

$$S_{x_j x_j} = \sum_i (x_{ij} - \bar{x}_j)^2$$

then

$$\text{Var}(\hat{\beta}_j) = \sigma^2 \left(\frac{1}{1 - R_j^2} \right) \frac{1}{S_{x_j x_j}}$$

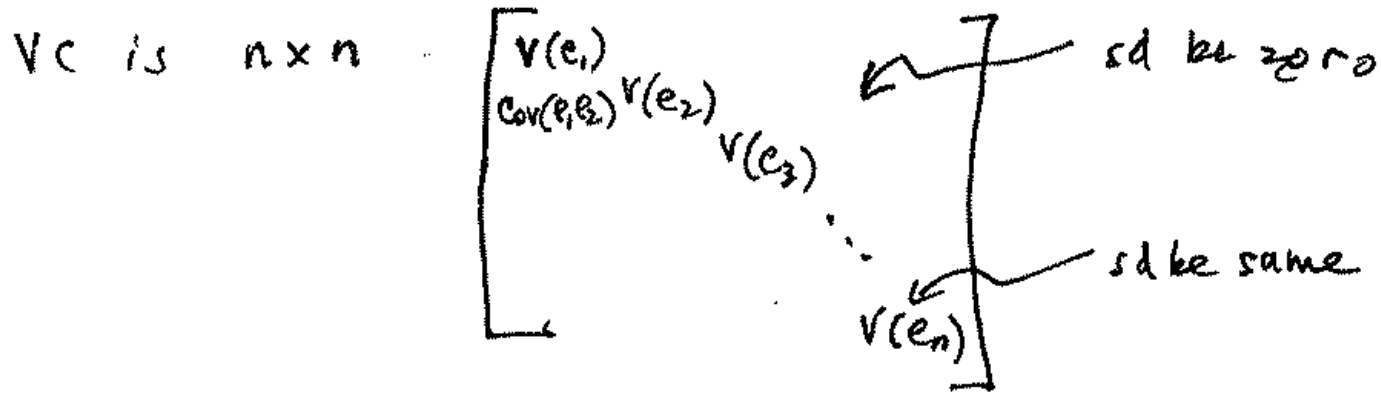
- So collinearity's main consequence is to reduce the efficiency of our estimates of β
- So if x_j does not vary much, then $\text{Var}(\hat{\beta}_j)$ will be large – and we can maximize $S_{x_j x_j}$ by spreading X as much as possible
- We call this factor $\frac{1}{1-R_j^2}$ a variance inflaction factor (the `faraway` package for R has a function called `vif()` you can use to compute it)
- *Orthogonality* means that variance is minimized when $R_j^2 = 0$

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 estiimte 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



- Consequences
  - Efficiency loss
  - Inconsistency: can no longer trust β_{OLS}
  - β_{OLS} is no longer the maximum likelihood estimator

Heteroskedasticity

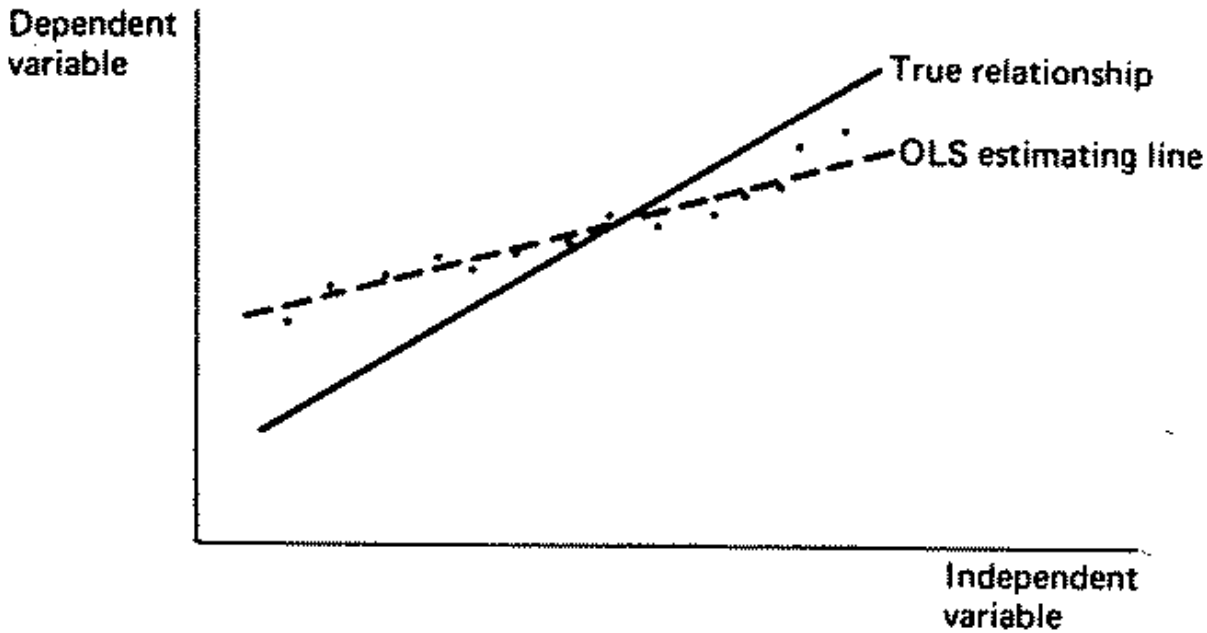
- Graphical inspection of residuals is best check
- Some tests also exist: Breusch-Pagan test, Goldfeld-Quandt test (see the `lmtest` library, it contains all of these)
- One solution is to use generalized least squares (GLS)
- A fix: White's heteroskedasticity-corrected standard errors:

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

- Can implement these very easily using the `vce(robust)` option to a `regress` command

Autocorrelated disturbances

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

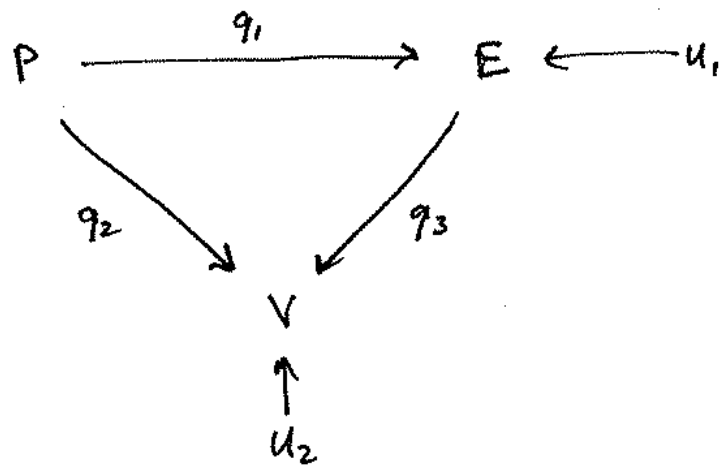


Autocorrelated disturbances

- 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 thee first order autocorrelation coefficient ρ
  - $0 < d < 4$
  - $d = 2$ when $\rho = 0$; if $d < 1.0$ then need correction
  - use `estat dwatson` (but the data must be `tsset <timevar>` first using a time variable
- Not all time-series models are complex, but this is an advanced topic with many forms of different models

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)
- Path model:



Simultaneous Equations: The Problem

- Question: What is the *total effect* of party ID on voting behaviour? includes:
  - the direct effect q_2, *plus*
  - the indirect effect $q_1 q_3$
- "Path analysis": uses correlations instead of covariances, so that all estimated relationships are standardized coefficients
- Substitute (1) into (2):

$$\begin{aligned} V &= q_2 P + q_3(q_1 P + U_1) + U_2 \\ &= q_2 P + q_3 q_1 P + q_3 U_1 + U_2 \\ &= (q_2 + q_3 q_1) P + q_3 U_1 + U_2 \\ &= \pi_i P + \nu \end{aligned}$$

General Multiequation Model

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

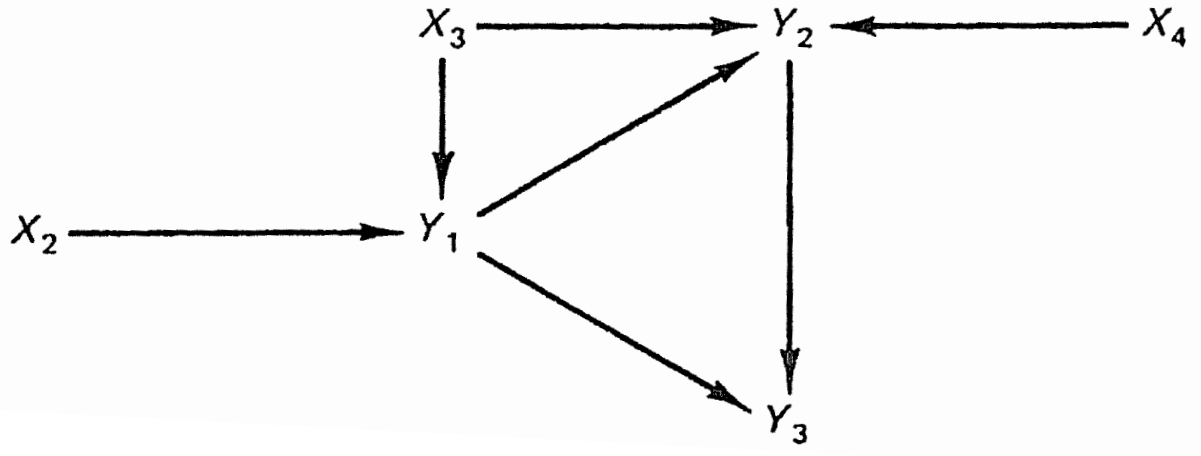
$$Y_2 = \gamma_{12}Y_1 + \beta_{12}X_1 + \beta_{22}X_2 + \beta_{32}X_3 + \beta_{42}X_4 + U_2$$

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

- where Y_m are endogenous variables, X_k are 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 total effect on the endogenous variables of a change in X, we must include all indirect effects through γ and β
- We do this through the reduced form equations

Example

Consider the following hierarchical model, where the error terms are correlated:



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

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

$$Y_3 = \gamma_{13}Y_1 + \gamma_{23}Y_2 + \beta_{13}X_1 + U_3, \quad (8.21)$$

Example cont.

The reduced form expressions for each endogenous variable are:

$$Y_1 = \beta_{11}X_1 + \beta_{21}X_2 + \beta_{31}X_3 + U_1 = \pi_{11}X_1 + \pi_{21}X_2 + \pi_{31}X_3 + V_1, \tag{8.22}$$

$$\begin{aligned} Y_2 &= (\beta_{12} + \gamma_{12}\beta_{11})X_1 + \gamma_{12}\beta_{21}X_2 + (\beta_{32} + \gamma_{12}\beta_{31})X_3 \\ &\quad + \beta_{42}X_4 + \gamma_{12}U_1 + U_2 \\ &= \pi_{12}X_1 + \pi_{22}X_2 + \pi_{32}X_3 + \pi_{42}X_4 + V_2, \end{aligned} \tag{8.23}$$

$$\begin{aligned} Y_3 &= (\beta_{13} + \gamma_{13}\beta_{11} + \gamma_{23}\beta_{12} + \gamma_{23}\gamma_{12}\beta_{11})X_1 + (\gamma_{13}\beta_{21} + \gamma_{23}\gamma_{12}\beta_{21})X_2 \\ &\quad + (\gamma_{13}\beta_{31} + \gamma_{23}\beta_{32} + \gamma_{23}\gamma_{12}\beta_{31})X_3 + \gamma_{23}\beta_{42}X_4 + (\gamma_{13} + \gamma_{23}\gamma_{12})U_1 \\ &\quad + \gamma_{23}U_2 + U_3 \\ &= \pi_{13}X_1 + \pi_{23}X_2 + \pi_{33}X_3 + \pi_{43}X_4 + V_3. \end{aligned} \tag{8.24}$$

Using reduced form equations to gauge total effect

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

	β_{23}	direct effect
	$\gamma_{23}\beta_{22}$	indirect effect from direct change in Y_2
	$\gamma_{13}\beta_{21}$	indirect effect from a direct change in Y_1
$+$	$\gamma_{23}\gamma_{12}\beta_{21}$	indirect effect due to changes in Y_2 caused by changes in Y_1
$=$	π_{23}	total effect

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 moels 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

Solution 1: Indirect Least Squares

- We directly estimate β_{11}, β_{21}, β_{31} from the regression on Y_1 (8.22)
- This leaves 3 "unknowns": γ_{12}, β_{12}, β_{32}

$$\pi_{22} = \gamma_{12}\beta_{21}$$
$$\pi_{12} = \beta_{12} + \gamma_{12}\beta_{11}$$
$$\pi_{32} = \beta_{32} + \gamma_{12}\beta_{31}$$

- We can estimate γ_{12} using the previous estimate for β_{21}
- and then we can use β_{11}, γ_{12} to estimate β_{12}
- and use β_{31}, γ_{12} to estimate β_{31}
- This method is known as indirect least squares

Solution 1: Indirect Least Squares cont.

- 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

Solution 2: Instrumental Variables

- 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)