Lurking variable.

A variable that has an important effect and yet is not included amongst the predictor variables under consideration.

Perhaps its existence is unknown or its effect unsuspected.

Mosteller and Tukey’s World War II example

Simpson’s Paradox.

Medical treatment and outcome

		outcome	
		success	failure
treatment	1	100	100
	2	110	80

success rate under 1: 100/200 = .50

success rate under 2: 110/190 = .579

treatment 2 looks the better

Actually the data were aggregated (collapsed) over gender

male

		success	failure
treatment	1	60	20
	2	100	50

success rate under 1: 60/80 = .75

success rate under 2: 100/150 = .67

treatment 1 looks the better

female

		success	failure
treatment	1	40	80
	2	10	30

success rate under 1: 40/120 = .33

success rate under 2: 10/40 = .25

treatment 1 looks the better

The conclusion of the study has been reversed.

The two sexes were weighted differently with treatment 1 going to 80 males and 120 females, while treatment 2 went to 150 males and 40 females.

Gender is a lurking variable

Difficulty results from a lurking variable and combination of unequal group sizes.

If all groups of same size, circumstance doesn’t arise.

Solutions for lurking variables – eliminate them, hold them constant, or make them part of the study.

Smoking and 20yr survival rate for 1314 English women. (started 1972-4)

		outcome	
		dead	alive
smoker	yes	139	443
	no	230	502

24% (139/582) of the smokers died and 31% (230/732) of the nonsmokers.

Better to be a smoker?

Paradox goes away if the data are broken down by age.

Smokers are more likely to die in all but one age group.

Statistics 215a – 11/12/03 – D.R. Brillinger

Exploratory time series analysis

Time series decomposition à la Cleveland.

Data y_t , t=1,2,…,T, with t *time*

Decomposition

response = *fit* + *residual*

Assuming the fit “smooth”, it can be estimated by loess(), polynomials, splines, …, by smoothers

Example. Monthly water usage for London, Ontario using log transform

$\text{data}, \text{fit}_1, \text{residual}_1$

A pattern remains!

$\text{data}_2 = \text{data} - \text{fit}_1$

$\text{data}_2, \text{fit}_2, \text{residual}_2$

stem-and-leaf(residual_2)

Smoothing and reroughing.

smooth(x,twice=T)

Robustly smooths a time series by means of running medians.

Running median can produce a jagged sequence.

Twicing. The process of smoothing, computing the residuals from the smooth, smoothing these and adding the two smoothed series together.

Hanning.

$$y_t = (y_{t-1} + 2y_t + y_{t+1})/4$$

Statistics 215a – 11/12/03 – D.R. Brillinger

Exploratory time series analysis.

Differencing.

$$\Delta y_t = y_t - y_{t-1}$$

Can lead to less of a pattern if the series is meandering or has a trend

Example. Closing prices of IBM

hist()

Phase portrait

plot y_t versus y_{t-1}

Autoregresion.

Approx y_t by linear function of $y_{t-1},\ldots,y_{t-p}$

Example 1. For IBM via OLS

$y_t \approx 1.286 + .982\ y_{t-1}$

$r^2 = .983$

suggesting consideration of difference

$$\Delta y_t \approx -0.273 + 0.084\ \Delta y_{t-1}$$

$r^2 = .007$

Example 2. London, Ontario water usage

autoregression with 1 lag

$$y_t \approx .383 + .920\ y_{t-1}$$

$r^2 = .854$

residual plot

structure remains

autoregression with lags 1 and 12

$$y_t \approx .030 + .347\ y_{t-1} + .651\ y_{t-12}$$

$R^2 = .919$

Could use robust least squares

The form of the matrix X^TX is worth noting in the time series case.

The problem is to minimize

$$\Sigma_t [y_t - \alpha - \beta_1 y_{t-1} - \ldots - \beta_p y_{t-p}]^2$$

a is such that

$$\Sigma_t [y_t - a - b_1 y_{t-1} - \ldots - b_p y_{t-p}] = 0$$

The normal equations for **b** contain

$$\Sigma_t [y_{t-i} - \bar{y}][y_{t-j} - \bar{y}]/T$$

in row i, column j of $X^T X$.

This is a Toeplitz matrix.

The normal equations are sometimes called the Yule-Walker equations.