In-class exercises: Problems in Regression Analysis

See the "Overview of Regression" supplement for more complete definitions of the three problems that arise in regression analysis.

I. Multicollinearity

Idea: Explanatory variables are highly correlated, so that they are individually statistically insignificant, but jointly significant.

Diagnosis: Low t-statistics, high F-statistic (and R^2).

Consequences: Difficulty with determining which of the explanatory variables is important. LS coefficients are unbiased and efficient. You are still able to conduct statistical inference on the coefficients.

Example workfile: Fish.wf1

Research question: How did the Pope's 1966 announcement that permitted Catholics to consume meat on non-Lent Fridays affect the consumption of fish?

$$\underset{\text{Economic Model: }}{F} = (\underset{-}{PF}, \underset{+}{PB}, \underset{+}{Yd}, \underset{+}{N})$$

where

F = pounds of fish consumed
PF = price of fish
PB = price of beef
Yd = disposable income
N = population of Catholics

Data: Time series data from the U.S., 1946-1970

Regression specification:

$$F_t = \alpha + \beta_1 PF_t + \beta_2 PB_t + \beta_3 \ln(Yd_t) + \beta_4 N_t + \beta_5 P_t + \varepsilon_t$$

where P_t is a dummy variable that takes on a value of 1 in years beginning in 1966.

Solution: Drop N and redefine a relative price of fish as RP = PF/PB to take care of the high correlation between PF and PB.

II. Heteroskedasticity

Idea: Error variance is not homoskedastic (the same) across observations.

Diagnosis: Residuals have unequal variance – plot the squared residuals. Conduct a White heteroskedasticity test.

Consequences: LS estimates of the standard error are incorrect, so we are unable to conduct statistical inference on the coefficient estimates.

Example workfile: Gas.wf1

Research question: How do gasoline taxes across states affect petroleum consumption across states? The

Department of Energy attempted to allocate gasoline to regions and states in the mid-1970s based on a model of the determinants of gasoline consumption.

$$\underset{\text{Economic Model: }}{PCON} = (\underset{-}{REG}, \underset{+}{TAX}, \underset{-}{})$$

where
PCON = petroleum consumption
REG = # of motor vehicle registrations
TAX= gasoline tax rate

Data: Cross section of states in the U.S.

Regression specification:

$$PCON_i = \alpha + \beta_1 REG_i + \beta_2 TAX_i + \varepsilon_i$$

Solution: Scale PCON and REG by state population (POP)

III. Autocorrelation
Idea: Errors are autocorrelated over time (residual in period t is correlated with the previous period's residual)

Diagnosis: Plot residuals over time, check correlogram, check Durbin-Watson statistic

Consequences: LS estimates of the standard error are incorrect, so we are unable to conduct statistical inference on the coefficient estimates.

Example workfile: Defend.wf1

Research question: Is defense spending in the Soviet Union a function of U.S. defense spending?

$$\underset{\text{Economic Model: }}{SD} = (\underset{+}{USD}, \underset{+}{SY}, \underset{+}{SP})$$

where
SD = Soviet defense spending
SY = Soviet GNP
SP= ratio of the number of USSR nuclear warheads to U.S. nuclear warheads = NR/NU

Data: Time series for the Soviet Union, 1960-1984; the CIA reports a high (SDH) and low (SDL) defense spending estimate for the Soviet Union

Regression specification: (SDH or SDL)

$$\ln(SDH_t) = \alpha + \beta_1 \ln(SY_t) + \beta_2 SP_t + \varepsilon_t$$

Solution: Estimate equation in differences, include a lagged dependent variable