Environmental Econometrics

Syngjoo Choi

Fall 2008

Syllabus I

- This is an introductory econometrics course which
  - assumes no prior knowledge on econometrics;
  - focuses on both theoretical results and practical uses of econometrics in (environmental) economic problems;
  - teaches a statistical software, STATA, in tutorial classes.
- Lecture and tutorial timetables
  - Lectures: every Monday (starting from Oct. 6 and ending on Dec. 8) , 9~11 am, B03 (Drayton House).
  - Tutorials: every Tuesday (starting from Oct. 7 and ending on Dec. 9), 9~11 am, B17 (computer room).
  - The tutorial classes will be given by TA, Jelmer Ypma (j.ypma@ucl.ac.uk).
  - Office hour: Monday, 3~4 pm and by appointment.

Syllabus II

- Main Textbook
  - J. Wooldridge (2008), *Introductory Econometrics: A Modern Approach*, 4th Ed., South-Western.
- Course materials
  - Lecture notes and exercises are available in my teaching webpage, http://www.homepages.ucl.ac.uk/~uctpsc0/Teaching.html.
  - Sample data for STATA exercises are also available.
  - The final exam and its answer keys from previous years are available.

Course Outline I

1. Linear regression models - *Wooldridge Ch. 2~5 and 7*
    - simple regression to multiple regression, ordinary least squares (OLS) estimation and goodness of fit.
    - hypothesis testing and large sample properties of OLS
2. Heteroskedasticity and Autocorrelation - *Wooldridge Ch. 8, 10 and 12*
    - consequences of heteroskedasticity and autocorrelation
    - testing for heteroskedasticity and autocorrelation
    - generalized least squares (GLS) estimation

Course Outline II

3. IV estimation and simultaneous equations models - *Wooldridge Ch. 15 and 16*
   - endogeneity, instrumental variables (IV) estimation and two-stage least squares
   - simultaneity bias, identification and estimation of simultaneous equations models
4. Limited dependent variable models - *Wooldridge Ch 17*
   - problems of using OLS for binary response models
   - maximum likelihood estimation, logit and probit models
   - ordered probit model, poisson regression model
   - censored dependent variables and Tobit models
5. Some simple panel data analysis - *Wooldridge Ch 13*
6. Time series analysis - *Wooldridge Ch. 12 and 18*
   - stationarity and nonstationarity; AR and MA processes; unit root
   - VAR; Granger causality

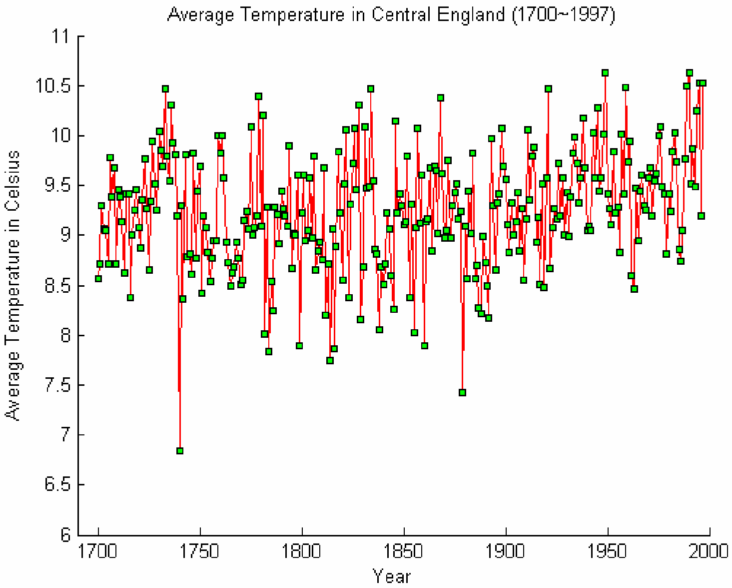
What is Econometrics?

- Statistical tools applied to economic problems
  - estimate economic relationships;
  - test economic theories modeling the causality of social and economic phenomena;
  - evaluate the impact and effectiveness of a given policy;
  - forecast the impact of future policies.
- It aims at providing not only a *qualitative* but also a *quantitative* answer.

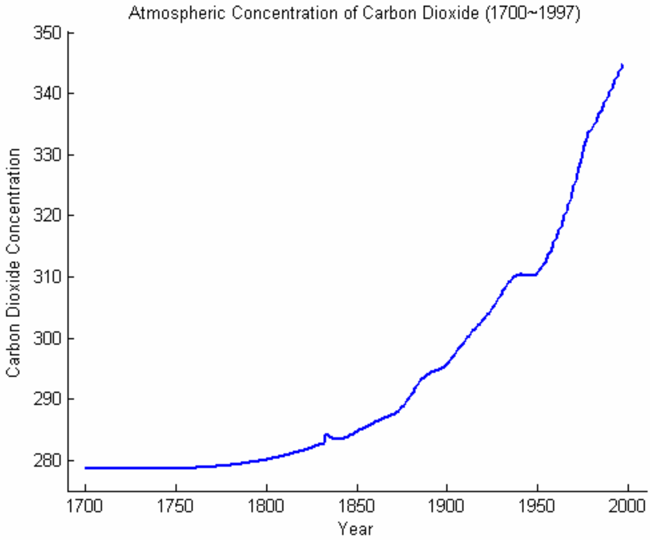
Example 1: Global Warming

- Measuring the extent of global warming
  - When did it start? How large is the effect?
  - Has it increased more in the last 50 years?
- What are the causes of global warming?
  - Does carbon dioxide cause global warming?
  - Are there any other determinants?
- What would be average temerature if carbon dioxide concentration is reduced by 10%?

Average Temperature in Central England (1700~1997)



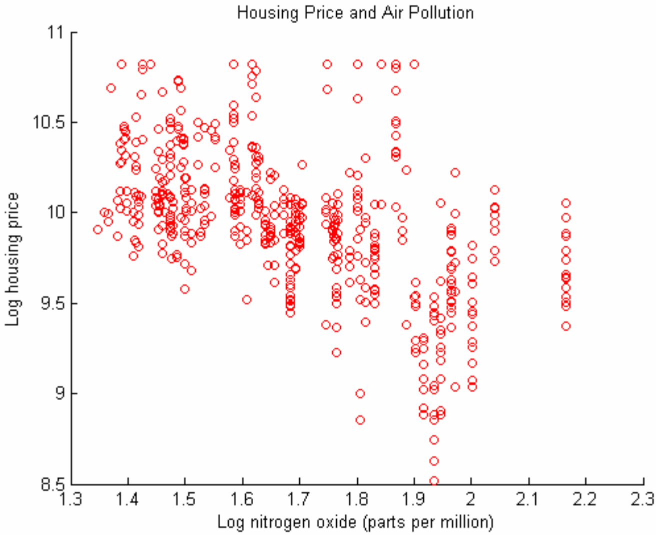
Atmospheric Concentration of Carbon Dioxide (1700~1997)



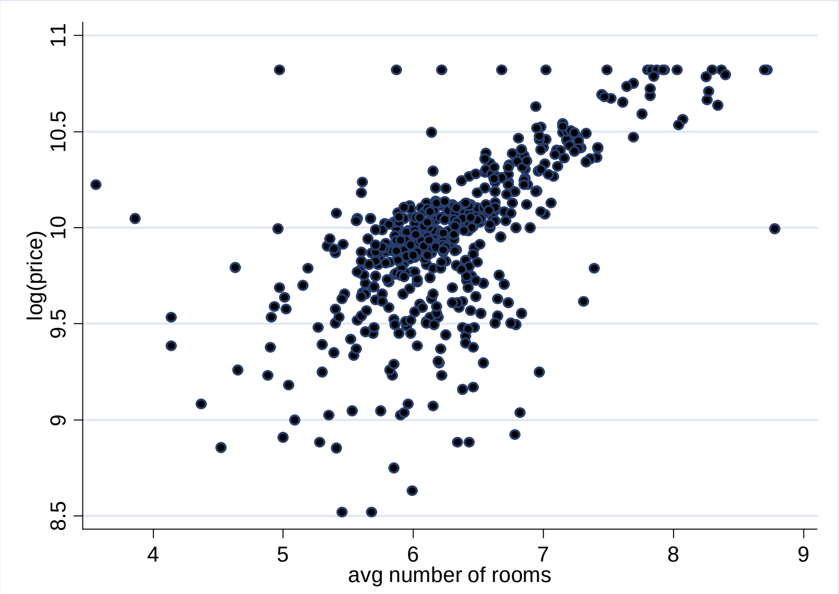
Example 2: Housing Prices and Air Pollution

- Measuring the effect of air pollution on housing prices
  - Does air pollution matter in determining housing prices?
  - If so, how much?
- Are there other determinants?
  - physical features of houses (e.g., number of rooms)
  - distance from workplaces
  - the quality of education in community

Median Housing Prices and Nitrogen Oxide (A sample of 506 communites in the Boston Area)



Median Housing Prices and Room Numbers



Causality

- We often observe that two variables are *correlated*.
  - Higher education leads higher income.
  - Individual smoking is related to peer smoking.
- If Y is *causally related* to X, then chaning X will lead to a change in Y.
- Correlation may not be due to causal relationships.
  - Some common factor may affect both variables.

Causality and Ceteris Paribus

- The notion of *ceteris paribus* (holding other variables constant) plays an important role in *causal analysis*.
  - Holding innate ability constant, how much does an increase in education increase in income?
  - Holding the individual taste of smoking, how much does an increase in peer smoking increase in individual smoking?
- This course will introduce how to deal with the issue of causality and ways of doing causal analysis.

The Simple Regression Model

- The simplest form in the regression model is the *two variable linear regression* model, called the *Simple Linear Regression Model*.

$$Y_i = \alpha + \beta X_i + u_i,$$

  - Y_i: dependent variable (explained variable; regressand)
  - X_i: independent variable (explanatory variable; regressor)
  - u_i: error term
  - $i = 1, ..., N$: the number of observation
- The error term or disturbance, u, represents all other factors affecting Y other than X.

Examples

$$Temp_i = \alpha + \beta Year_i + u_i,$$

$$Hprice_i = \alpha + \beta Nox_i + u_i.$$

- X has a linear effect on Y if all other factors are held constant.

$$\Delta Y = \beta \Delta X \quad \text{if} \quad \Delta u = 0.$$

  - The linearity implies that a one-unit change in X has the *same* effect on Y, *regardless of the intial value of* X.
- The slope parameter in the relationship between Y and X is meant to capture the effect of X on Y.
- In order to interpret so, we need to make an assumption regarding how u and X are **UNrelated**.

Assumption 1

- Assumption (Zero Conditional Mean)

$$E(u|X) = 0.$$

It says that for any given value of X, the average of the error term u is equal to 0.

Assumption 1

- Assumption (Zero Conditional Mean)

$$E\,(u|X) = 0.$$

  It says that for any given value of X, the average of the error term u is equal to 0.

- Thus, observing a high or a low value of X does not imply a high or a low value of u. That is, X and u are *uncorrelated*.

Assumption 1

- Assumption (Zero Conditional Mean)

$$E(u|X) = 0.$$

It says that for any given value of X, the average of the error term u is equal to 0.

- Thus, observing a high or a low value of X does not imply a high or a low value of u. That is, X and u are *uncorrelated*.
- Example - wage equation

Assumption 1

- Assumption (Zero Conditional Mean)

$$E\left(u|X\right) = 0.$$

  It says that for any given value of X, the average of the error term u is equal to 0.
- Thus, observing a high or a low value of X does not imply a high or a low value of u. That is, X and u are *uncorrelated*.
- Example - wage equation
  - Assume that u only reflects innate ability.

$$E\left(u|years\ of\ education\right) = 0$$

Assumption 1

- Assumption (Zero Conditional Mean)

$$E(u|X) = 0.$$

  It says that for any given value of X, the average of the error term u is equal to 0.
- Thus, observing a high or a low value of X does not imply a high or a low value of u. That is, X and u are *uncorrelated*.
- Example - wage equation
  - Assume that u only reflects innate ability.

$$E(u|years\ of\ education) = 0$$

  - The zero-conditional-mean assumption requires that the average level of ability is the same regardless of years of education. Is it reasonable?

An Implication of Zero Conditional Mean Assumption

- The zero conditional mean assumption implies that the population regression function, $E\left(Y|X\right)$, is a linear function of X.

$$E\left(Y|X\right) = \alpha + \beta X$$

- For any given value of X, the distribution of Y is centered around $E\left(Y|X\right)$. (figure)

Regression Problem - Least Squares

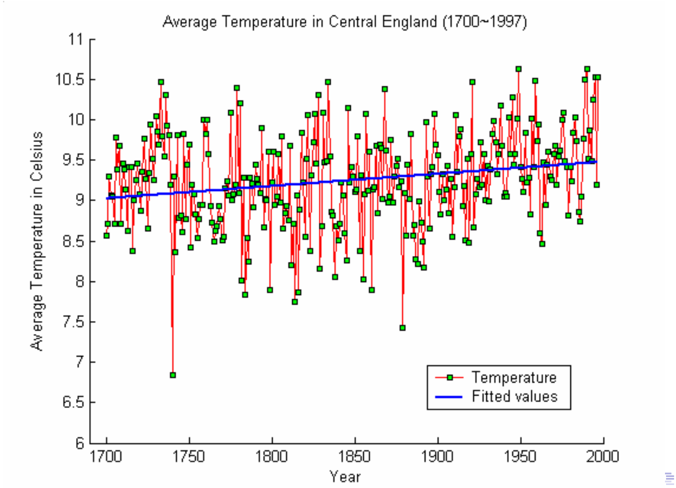
- Consider a set of data $\{(X_i, Y_i)\}_{i=1}^{N}$ and we want to obtain estimates of the intercept and slope.
- The most popular method in econometrics is the *least squares estimation*.
  - choose $\widehat{\alpha}$ and $\widehat{\beta}$ to minimize the sum of squared residuals

$$\sum_{i=1}^{N} \widehat{u}_i^2 = \sum_{i=1}^{N} \left(Y_i - \widehat{\alpha} - \widehat{\beta} X_i \right)^2 .$$

- The estimates given from this minimization problem is called the *ordinary least squares* (OLS).
- Using the OLS estimation, we have the estimated regression line for the unknown population regression function (figure).

- The OLS estimated regression line is given by

$$Temp_i = 6.45 + 0.0015 \times Year_i.$$



Model Specification

- Linear model

$$Y_i = \alpha + \beta X_i + u_i$$

  - When X goes up by 1 unit, Y goes by β units.

- Log-log model (constant elasticity model)

$$\ln Y_i = \alpha + \beta \ln X_i + u_i$$

  - When X goes up by 1%, Y goes up by β%.

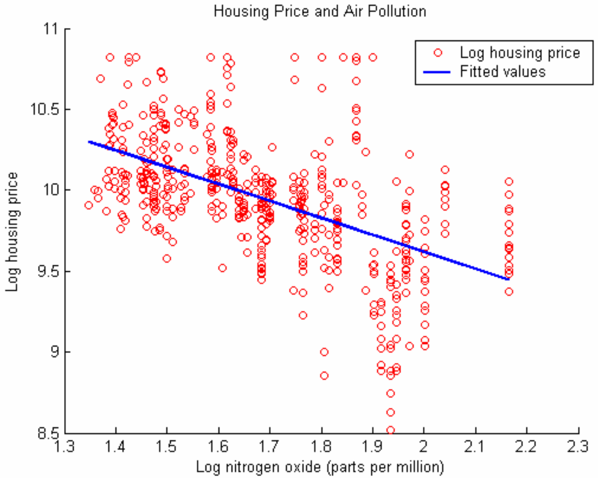
- Log-linear model

$$\ln Y_i = \alpha + \beta X_i + u_i$$

  - When X goes up by 1 unit, Y goes up by 100β%.

- The estimated regression line is given

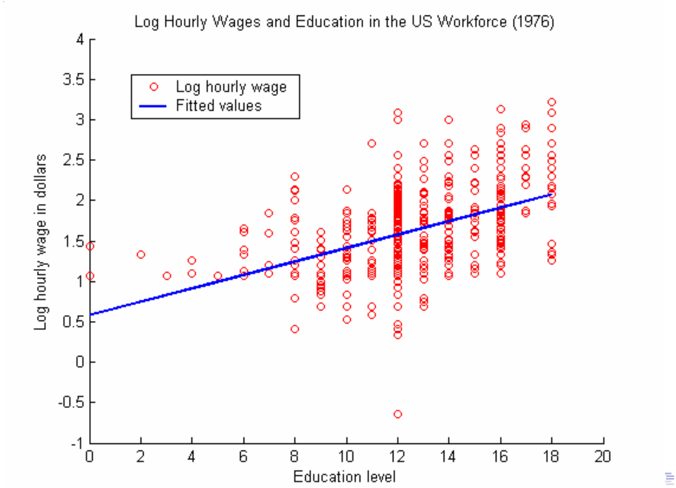
$$\ln Hprice_i = 11.71 - 1.04 \times \ln Nox_i.$$



Example 2: Wage Equation

- The estimated regression line is given

$$\ln Wage_i = 0.58 + 0.0083 \times Educ_i.$$



The Structure of Data

- Times Series Data
  - Data on variables observed time. Examples include stock prices, consumer price index, annual homicide rates, GDP, and temperature changes cross time.
- Cross Section Data
  - Data at a given point in time on individuals, households or firms. Examples are data on expenditures, income and employment (say, in 1999).
- Panel or Longitudinal Data
  - Data on a time series for each cross-sectional member.

Type of Variables

- Continuous
  - temperature; wage; housing prices.
- Categorical/Qualitative
  - ordered
    - years of schoolding; survey answers such that small/medium/large.
  - unordered
    - decisions such as Yes/No; gender(male/female).
- The course will explain later how to deal with qualitative dependent variables.

Properties of OLS

First-order Conditions

- The least squares problem, as a reminder, is

$$\min_{\widehat{\alpha},\widehat{\beta}} \sum_{i=1}^{N} \widehat{u}_i^2 = \sum_{i=1}^{N} \left(Y_i - \widehat{\alpha} - \widehat{\beta} X_i \right)^2,$$

- The first-order conditions (FOC) are given by

$$\frac{\partial \sum_{i=1}^{N} \widehat{u}_i^2}{\partial \widehat{\alpha}} = -2 \sum_{i=1}^{N} \left(Y_i - \widehat{\alpha} - \widehat{\beta} X_i \right) = 0,$$

$$\frac{\partial \sum_{i=1}^{N} \widehat{u}_i^2}{\partial \widehat{\beta}} = -2 \sum_{i=1}^{N} \left(Y_i - \widehat{\alpha} - \widehat{\beta} X_i \right) X_i = 0.$$

The OLS Estimator

- Solving the two FOC equations, we have the OLS estimators for the intercept and the slope parameters:

$$\widehat{\alpha} = \overline{Y} - \widehat{\beta}\overline{X},$$

$$\widehat{\beta} = \frac{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2},$$

  where $\overline{Z} = \sum_{i=1}^{N} Z_i / N$. We need a condition that $\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2 > 0$.
- The estimate of the slope coefficient is simply the sample covariance between X and Y divided by the smaple variance of X.
- (Diagression) An *estimator* is a random variable and an *estimate* is a realization of an estimator.

Algebraic Properties

- Property 1: The sum of OLS residuals in zero.

$$\sum_{i=1}^{N} \widehat{u}_i = \sum_{i=1}^{N} \left(Y_i - \widehat{\alpha} - \widehat{\beta} X_i \right) = 0.$$

- Property 2: The sample covariance between the independent variable X and the OLS residual $\widehat{u}$ is zero.

$$\sum_{i=1}^{N} X_i \widehat{u}_i = \sum_{i=1}^{N} X_i \left(Y_i - \widehat{\alpha} - \widehat{\beta} X_i \right) = 0.$$

- Property 3: The OLS estimates decompose each Y_i into a fitted value $\widehat{Y}_i$ and a residual $\widehat{u}_i$.

$$Y_i = \widehat{Y}_i + \widehat{u}_i \Longrightarrow \overline{Y} = \overline{\widehat{Y}}.$$

Goodness of Fit I

- We want to measure how well the model fits the data.
- The *R-squared* of the regression is defined as the ratio of the explained sum of squares to the total sum of squares.
  - Total sum of squares (TSS): $TSS = \sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2$.
  - Explained sum of squares (ESS)

$$ESS = \sum_{i=1}^{N} \left(\widehat{Y}_i - \overline{Y}\right)^2 = \sum_{i=1}^{N} \left[\widehat{\beta}\left(X_i - \overline{X}\right)\right]^2.$$

  - Residual sum of squares (RSS): $RSS = \sum_{i=1}^{N} \widehat{u}_i^2$.
  - The R-squared of the regression is

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}.$$

Goodness of Fit II

- The R-squared is a measure of how much of the variance of Y is explained by th regressor X.
- The value of R-squared is always between 0 and 1. If R-squared is equal to 1, then OLS provides a perfect fit to the data.
- A low R-squared is not necessarily an indication that the model is wrong. It is simply that the regressor has low explanatory power.

An Example: Housing Prices and Air Pollution

Variable	Coefficient
ln *Nox*	-1.043
constant	11.71
Model sum of squares	22.29
Residual sum of squares	62.29
Total sum of Squares	84.58
R-squared	0.26
Number of observation	506

Statistical Properties of OLS

- Given a specific sample of data $\{(X_i, Y_i)\}_{i=1}^{N}$, $\widehat{\alpha}$ and $\widehat{\beta}$ are realized values of the OLS estimator in the simple linear regression model.
- It means that if we have a different sample from the same population, then we may have different values of the slope and intercept estimates.
- We want the estimators to have desirable properties:
    - **Unbiasedness**
    - **Efficiency**

Assumptions on the Simple Linear Regression Model

- Assumption 1: Zero Conditional Mean

$$E\left(u_i|X\right) = 0.$$

- Assumption 2: Homoskedasticity

$$Var\left(u_i|X\right) = E\left[u_i - E\left(u_i|X\right)|X\right]^2 = \sigma^2.$$

- Assumption 3: No correlation among error terms

$$Cov\left(u_i, u_j|X\right) = 0, \forall i \neq j.$$

- Assumption 4: Sufficient variation in X

$$Var\left(X\right) > 0.$$

Unbiasedness

- Definition: Estimators $\widehat{\alpha}$ and $\widehat{\beta}$ are unbiased if

$$E\left(\widehat{\alpha}\right) = \alpha \text{ and } E\left(\widehat{\beta}\right) = \beta.$$

- Unbiasedness does **NOT** mean that the estimate we get with a particular sample is equal to the true value.
- If we could *indefinitely* draw random samples of the same size N from the population, compute an estimate each time, and then average these estimates over all random samples, we would obtain the true value.

An Example

- Suppose the true model is

$$Y_i = 1 + 2X_i + u_i,\ u_i \sim iid\ N(0,1).$$

- We generate a set of random samples, each of which contains 14 observations.

	$\widehat{\alpha}$	$\widehat{\beta}$
Random Sample 1	1.2185099	1.5841877
Random Sample 2	0.8250200	2.5563998
Random Sample 3	1.3752522	1.3256603
Random Sample 4	0.9216356	2.1068873
Random Sample 5	1.0566855	2.1198698
Random Sample 6	1.0482750	1.8185249
Random Sample 7	0.9140797	1.6573014
Random Sample 8	0.7885023	2.9571939
Random Sample 9	0.6581880	2.2935987
Random Sample 10	1.0852489	2.3455551
Average across 10 random samples	0.9891397	2.0765179
Average across 500 random samples	0.9899374	2.0049863

Unbiasedness of OLS Estimator

- Recall that the OLS estimators of the intercept and the slope are

$$\widehat{\alpha} = \overline{Y} - \widehat{\beta}\overline{X}$$

and

$$\widehat{\beta} = \frac{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2}.$$

- First note that within a sample

$$\overline{Y} = \alpha + \beta\overline{X} + \overline{u}.$$

Hence, for any $i = 1, ..., N$,

$$Y_i - \overline{Y} = \beta\left(X_i - \overline{X}\right) + u_i - \overline{u}.$$

- Substitute this in the expression for $\widehat{\beta}$:

$$\begin{aligned}\widehat{\beta} &= \frac{\sum_{i=1}^{N}\left[\beta\left(X_i-\overline{X}\right)^2+\left(X_i-\overline{X}\right)(u_i-\overline{u})\right]}{\sum_{i=1}^{N}\left(X_i-\overline{X}\right)^2} \\ &= \beta+\frac{\sum_{i=1}^{N}\left(X_i-\overline{X}\right)(u_i-\overline{u})}{\sum_{i=1}^{N}\left(X_i-\overline{X}\right)^2}.\end{aligned}$$

The second part of the right-hand side is called the sampling error. If the estimator is unbiased, then this error will have expected value zero.

-

$$\begin{aligned}\mathbf{E}\left(\widehat{\beta}|X\right) &= \beta+\mathbf{E}\left[\frac{\sum_{i=1}^{N}\left(X_i-\overline{X}\right)(u_i-\overline{u})}{\sum_{i=1}^{N}\left(X_i-\overline{X}\right)^2}|X\right] \\ &= \beta+\frac{\sum_{i=1}^{N}\left(X_i-\overline{X}\right)\mathbf{E}\left[(u_i-\overline{u})\,|X\right]}{\sum_{i=1}^{N}\left(X_i-\overline{X}\right)^2} \\ &= \beta, \qquad \text{using which assumption?}\end{aligned}$$

- Now,

$$\begin{aligned}\widehat{\alpha} &= \overline{Y} - \widehat{\beta}\overline{X} \\ &= \alpha + \left(\beta - \widehat{\beta}\right)\overline{X} + \overline{u}.\end{aligned}$$

Then,

$$\begin{aligned}\mathbf{E}\left(\widehat{\alpha}|X\right) &= \alpha + \mathbf{E}\left[\left(\beta - \widehat{\beta}\right)|X\right]\overline{X} + \mathbf{E}\left(\overline{u}|X\right) \\ &= \alpha.\end{aligned}$$

Variances of the OLS Estimators

- We have shown that the OLS estimator is unbiased under the assumptions.

- But how sensitive are the results to random changes to our sample? The variance of the estimators is a measure for this question.

- The definition of the variance is

$$Var\left(\widehat{\beta}|X\right) = \mathbf{E}\left[\left(\widehat{\beta} - \mathbf{E}\left(\widehat{\beta}\right)\right)^2 |X\right].$$

- Recall that

$$\mathbf{E}\left(\widehat{\beta}\right) = \beta$$

and

$$\widehat{\beta} - \beta = \frac{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)(u_i - \overline{u})}{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}.$$

- Thus,

$$
\begin{aligned}
Var\left(\widehat{\beta}|X\right) &= \mathbf{E}\left[\left(\frac{\sum_{i=1}^{N}\left(X_i-\overline{X}\right)(u_i-\overline{u})}{\sum_{i=1}^{N}\left(X_i-\overline{X}\right)^2}\right)^2|X\right] \\
&= \frac{1}{\left[\sum_{i=1}^{N}\left(X_i-\overline{X}\right)^2\right]^2} \\
&\quad \times \left[\begin{array}{c} \sum_{j=1}^{N}\sum_{i=1}^{N}\left(X_i-\overline{X}\right)\left(X_j-\overline{X}\right) \\ \times\mathbf{E}\left[(u_i-\overline{u})\left(u_j-\overline{u}\right)|X\right] \end{array}\right]
\end{aligned}
$$

- From Assumption 2 (homoskedasticity) and Assumption 3 (no autocorrelation),

$$\mathbf{E}\left[(u_i-\overline{u})^2\,|X\right]=\sigma^2$$

and

$$\mathbf{E}\left[\left((u_i-\overline{u})\left(u_j-\overline{u}\right)\right)|X\right]=0,\ \forall i\neq j.$$

- Then,

$$\begin{aligned} Var\left(\widehat{\beta}|X\right) &= \frac{\sigma^2}{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2} \\ &= \frac{1}{N}\frac{\sigma^2}{\widehat{Var\left(X\right)}}. \end{aligned}$$

- Properties of the variance of $\widehat{\beta}$

  - The variance increases with the variance of the error term, σ^2.

  - The variance decreases with the variance of X, $\widehat{Var\left(X\right)}$.

  - The variance decreases with the sample size, N.

  - The standard error is the square root of the variance:

$$SE\left(\widehat{\beta}|X\right) = \sqrt{Var\left(\widehat{\beta}|X\right)}.$$

- In practice, we do not know the variance of the error term, σ^2, which needs to be estimated. Using the residuals, $\widehat{u}_i$, we have an unbiased estimator of σ^2:

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{N} \widehat{u}_i^2}{N-2} = \frac{RSS}{N-2}.$$

An Example: Housing Prices and Air Pollution

Variable	Coefficient	Std. Err.
log *Nox*	−1.043	0.078
constant	11.71	0.132
R-squared		0.26
Number of observation		

Efficiency

- An estimator is *efficient* if given the assumptions we make, its variance is the smallest possible in the class of estimators we consider.

- We consider the class of linear unbiased estimators. An estimator is *linear* if and only if it can be expressed as a linear function of the data on the dependent variable.

- Note that the OLS estimator is a *linear* estimator:

$$\widehat{\beta} = \frac{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)}{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2} Y_i - \frac{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)\overline{Y}}{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}$$

- Consider another linear estimator of the slope. Define $Z_i = X_i^2$ and a slope estimator as

$$\widetilde{\beta} = \frac{\sum_{i=1}^{N}\left(Z_i - \overline{Z}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{N}\left(Z_i - \overline{Z}\right)\left(X_i - \overline{X}\right)}.$$

Then,

$$\mathbf{E}\left[\widetilde{\beta}|X\right] = \beta. \text{ (why?)}$$

- It can be also shown that

$$Var\left(\widehat{\beta}|X\right) \leq Var\left(\widetilde{\beta}|X\right).$$

- In fact, the OLS estimator has the smallest variance among the class of linear unbiased estimators, under the assumptions we made.

The Gauss Markov Theorem

- Given the assumptions we made, the OLS estimator is a **Best Linear Unbiased Estimator (BLUE)**.
- The means that the OLS estimator is the most *efficient (least variance)* estimator in the class of *linear unbiased* estimator.