Advanced Econometrics - Lecture 1

The Linear Regression Model: A Review

Econometrics …

- … consists of the application of statistical data and techniques to mathematical formulations of economic theory. It serves to test the hypotheses of economic theory and to estimate the implied interrelationships." (Tinbergen, 1952)
- … is the interaction of economic theory, observed data and statistical methods. It is the interaction of these three that makes econometrics interesting, challenging, and perhaps, difficult." (Verbeek, 2008)
- … is a methodological science with the elements
  - economic theory
  - mathematical language
  - statistical methods
  - software

The Contents

1. Review of linear regression and the OLS estimator
2. Heteroskedasticity and autocorrelation (MV, Ch.4)
3. Endogeneity, instrumental variables and GMM (MV, Ch.5)
4. Maximum likelihood estimation and specification tests (MV, Ch.7)
5. Univariate time series models (MV, Ch.8)
6. Multivariate time series models (MV, Ch.9)

Advanced Econometrics - Lecture 1

- Regression: a descriptive tool
- Economic models
- OLS estimation
- Properties of OLS estimators
- *t*-test and *F*-test
- Asymptotic properties of OLS estimators
- Multicollinearity
- Model specification and tests

The Linear Model

Y: explained variable
X: explanatory or regressor variable
The model describes the data-generating process of *Y* under the condition *X*

simple linear regression model

$$Y = \alpha + \beta X$$

β: coefficient of *X*
α: intercept

multiple linear regression model

$$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_K X_K$$

Fitting a Model to Data

Choice of values b_1, b_2 for model parameters β_1, β_2 of $Y = \beta_1 + \beta_2 X$, given the observations (y_i, x_i), $i = 1,\ldots,N$

Principle of (Ordinary) Least Squares or OLS:

$$b_i = \arg\min_{\beta 1,\ \beta 2} S(\beta_1, \beta_2),\ i=1,2$$

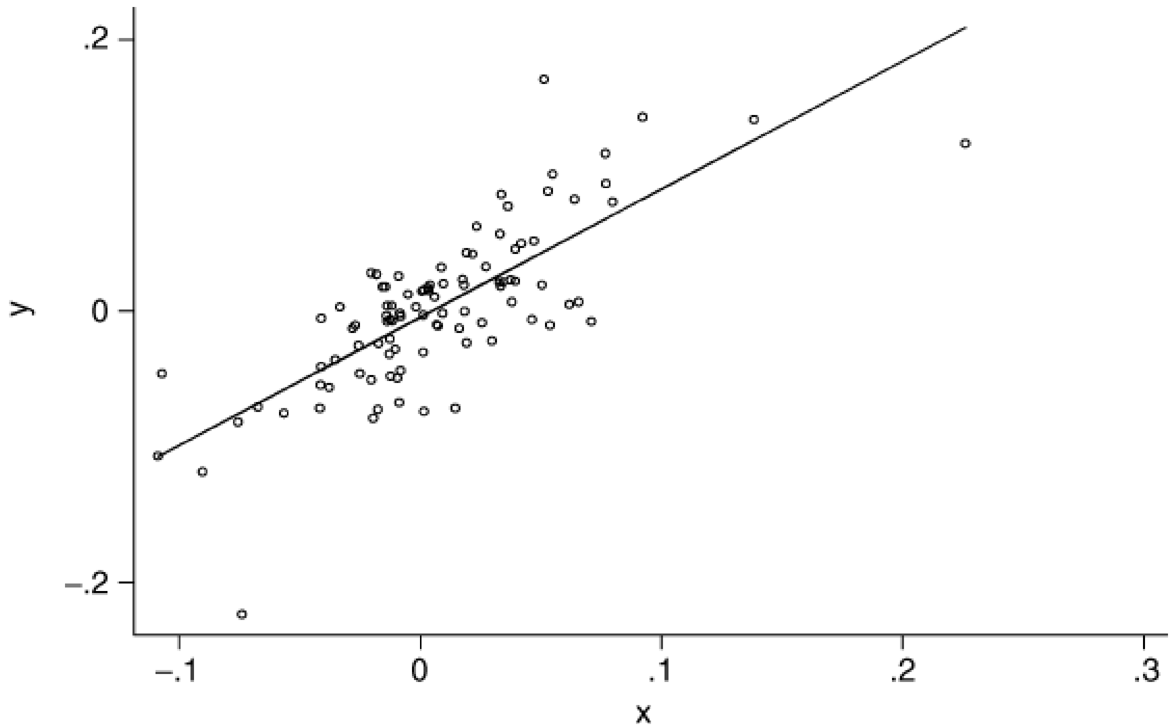
Objective function: sum of the squared deviations

$$S(\beta_1, \beta_2) = \Sigma_i [y_i - (\beta_1 + \beta_2 x_i)]^2 = \Sigma_i u_i^2$$

Deviations between observation and fitted values: $u_i = y_i - (\beta_1 + \beta_2 x_i)$

Observations and Fitted Regression Line

Simple linear regression: Fitted line and observation points (Verbeek, Figure 2.1)



OLS-Estimators

Equating the partial derivatives of $S(\beta_1, \beta_2)$ to zero: normal equations

$$b_1 + b_2 \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i$$

$$b_1 \sum_{i=1}^{N} x_i + b_2 \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i y_i$$

OLS-estimators b_1 und b_2 result in

$$\boxed{\begin{aligned} b_2 &= \frac{}{s_x^2} \\ b_1 &= \; - \; \end{aligned}}$$

with mean values $\bar{x}$ and $\bar{y}$

and second moments

$$s_{xy} = \frac{1}{N} \sum x_i - \bar{x})(y_i - \bar{y})$$

$$s_x^2 = \frac{1}{N} \sum x_i - \bar{x})^2$$

Example: Individual Wages

Sample (US National Longitudinal Survey, 1987): wage rate (per hour), gender, experience and years of schooling; N = 3294 individuals (1569 females)

Average wage rate (p.h.): 6.31$ for males, 5.15$ for females

Model (see eq. (2.39) in Verbeek):

$$wage_i = \beta_1 + \beta_2\, male_i + \varepsilon_i$$

$male_i$: male dummy, has value 1 if individual is male, otherwise value 0

OLS-estimation gives

$$wage_i = 5.15 + 1.17 * male_i$$

Compare with averages!

Example: Individ. Wages, cont'd

OLS estimated wage equation (Table 2.1, Verbeek)

Dependent variable: *wage*		
Variable	Estimate	Standard error
constant	5.1469	0.0812
male	1.1661	0.1122
$s = 3.2174$	$R^2 = 0.0317$	$F = 107.93$

$wage_i = 5.15 + 1.17 * male_i$

male: 6.313, female: 5.150

OLS-Estimators: The General Case

Model for Y contains K-1 explanatory variables

$$Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_K X_K = x'\beta$$

with $x = (1, X_2, \ldots, X_K)'$ and $\beta = (\beta_1, \beta_2, \ldots, \beta_K)'$

Observations: $(y_i, x_i) = (y_i, (1, x_{i2}, \ldots, x_{iK})')$, $i = 1, \ldots, N$

OLS-estimates $b = (b_1, b_2, \ldots, b_K)'$ are obtained by minimizing

$$S(\beta) = \sum_{i=1}^{N} (y_i - x_i'\beta)^2$$

this results in

$$b = \left(\sum_{i=1}^{N} x_i x_i'\right)^{-1} \sum_{i=1}^{N} x_i y_i$$

Best Linear Approximation

Given the observations: $(y_i, x_i') = (y_i, (1, x_{i2}, \ldots, x_{iK})')$, $i = 1, \ldots, N$

For y_i, the linear combination or fitted value

$$\hat{y}_i = x_i' b$$

is the best linear combination of Y from $X_2, \ldots, X_K$ and a constant (the intercept)

Residuals: $e_i = y_i - x_i' b$, $i = 1, \ldots, N$

- Minimum value of objective function: $S(b) = \Sigma_i e_i^2$
- Orthogonality of $e = (e_1, \ldots, e_N)'$ to each $x_i = (x_{1i}, \ldots, x_{Ni})'$: $e' x_i = 0$
- $\Sigma_i e_i = 0$: average residual is zero, if the model has an intercept

Matrix Notation

N observations

$(y_1, x_1), \ldots, (y_N, x_N)$

Model: $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, $i = 1, \ldots, N$, or

$$y = X\beta + \varepsilon$$

with

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix},\ X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix},\ \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix},\ \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{pmatrix}$$

OLS-estimates:

$$b = (X'X)^{-1}X'y$$

Advanced Econometrics - Lecture 1

- Regression: a descriptive tool
- Economic models
- OLS estimation
- Properties of OLS estimators
- *t*-test and *F*-test
- Asymptotic properties of OLS estimators
- Multicollinearity
- Model specification and tests

Economic Models

Describe economic relationships (not only a set of observations), have an economic interpretation

Linear regression model:

$$y_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_K x_{iK} + \varepsilon_i = x_i'\beta + \varepsilon_i$$

- Variables y_i, x_{i2}, …, x_{iK}: observable
- Error term ε_i (disturbance term) contains all influences that are not included explicitly in the model; unobservable; assumption $E\{\varepsilon_i \mid x_i\} = 0$ gives

  $$E\{y_i \mid x_i\} = x_i'\beta$$

  the model describes the expected value of *y* given *x*
- Unknown coefficients β_1, …, β_K: β_k measure the change of *Y* if X_k changes

Regression Coefficients

Linear regression model:

$$y_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_K x_{iK} + \varepsilon_i = x_i'\beta + \varepsilon_i$$

Coefficient β_k measures the change of Y if X_k changes by one unit and all other X values remain the same (ceteris paribus condition); marginal effect of changing X_k on Y

$$\frac{\partial E\{y_i | x_i\}}{\partial x_{ik}} = \beta_k$$

Example

- Wage equation: $wage_i = \beta_1 + \beta_2\, male_i + \beta_3\, school_i + \beta_4\, exper_i + \varepsilon_i$

  β_3 measures the impact of one additional year at school upon a person's wage, keeping gender and years of experience fixed

Regression Coefficients, cont’d

The marginal effect of a changing regressor may be non-constant

Example

- Wage equation: *wage*$_i$ = $\beta_1 + \beta_2$ *male*$_i$ + β_3 *age*$_i$ + β_4 *age*$_i^2$ + ε_i

  the impact of changing age (ceteris paribus) depends on age:

$$\frac{\partial E\{y_i | x_i\}}{\partial\, age_i} = \beta_3 + 2 age_i \beta_4$$

Elasticities

Elasticity: measures the *relative* change in the dependent variable *Y* due to a *relative* change in X_k

- For a linear regression, the elasticity of *Y* with respect to X_k is

$$\frac{\partial E\{y_i | x_i\} / E\{y_i | x_i\}}{\partial x_{ik} / x_{ik}} = \frac{x_{ik}}{x_i'\beta} \beta_k$$

- For a loglinear model log y_i = (log x_i)' β + ε_i, (log x_i)' = (1, log x_{i2},…, log x_{ik}), the elasticities are the coefficients β

$$\frac{\partial E\{y_i | x_i\} / E\{y_i | x_i\}}{\partial x_{ik} / x_{ik}} \approx \frac{\partial E\{\log y_i | \log x_i\}}{\partial \log x_{ik}} = \beta_k$$

Advanced Econometrics - Lecture 1

- Regression: a descriptive tool
- Economic models
- OLS estimation
- Properties of OLS estimators
- *t*-test and *F*-test
- Asymptotic properties of OLS estimators
- Multicollinearity
- Model specification and tests

Fitting Economic Models to Data

Observations allow

- to estimate parameters
- to assess how well the data-generating process is represented by the model, i.e., how well the model coincides with reality
- to improve the model if necessary

Fitting a linear regression model to data

- Parameter estimates $b = (b_1, \ldots, b_K)'$ for coefficients $\beta = (\beta_1, \ldots, \beta_K)'$
- Standard errors se(b_k) of the estimates b_k, k=1,…,K
- t-statistics, F-statistic, R^2, Durbin Watson test-statistic, etc.

OLS Estimator and OLS Estimates *b*

OLS estimates *b* are a realization of the OLS estimator

The OLS estimator is a random variable

- Observations are a random sample from the population of all possible samples
- Observations are generated by some random process

Distribution of the OLS estimator

- Actual distribution not known
- Theoretical distribution determined by assumptions on
  - model specification
  - the error term ε_i and regressor variables x_i

Quality criteria (bias, accuracy, efficiency) of OLS estimates are determined by the properties of the distribution

Gauss-Markov Assumptions

Observation y_i is a linear function

$$y_i = x_i'\beta + \varepsilon_i$$

of observations x_{ik}, k =1, …, K, of the regressor variables and the error term ε_i

for i = 1, …, N; $x_i' = (x_{i1}, \ldots, x_{iK})$; $X = (x_{ik})$

A1	$E\{\varepsilon_i\} = 0$ for all i
A2	all ε_i are independent of all x_i (exogeneous x_i)
A3	$V\{\varepsilon_i\} = \sigma^2$ for all i (homoskedasticity)
A4	$Cov\{\varepsilon_i, \varepsilon_j\} = 0$ for all i and j with $i \neq j$ (no autocorrelation)

Systematic Part of the Model

The systematic part $E\{y_i \mid x_i\}$ of the model $y_i = x_i'\beta + \varepsilon_i$, given observations x_i , is derived under the Gauss-Markov assumptions as follows:

(A2) implies $E\{\varepsilon \mid X\} = E\{\varepsilon\} = 0$ and $V\{\varepsilon \mid X\} = V\{\varepsilon\} = \sigma^2 I_N$

- Observations x_i do not affect the properties of ε
- The systematic part

  $$E\{y_i \mid x_i\} = x_i'\beta$$

  can be interpreted as the conditional expectation of y_i, given observations x_i

Advanced Econometrics - Lecture 1

- Regression: a descriptive tool
- Economic models
- OLS estimation
- Properties of OLS estimators
- *t*-test and *F*-test
- Asymptotic properties of OLS estimators
- Multicollinearity
- Model specification and tests

Is the OLS estimator a good estimator?

- Under the Gauss-Markov assumptions, the OLS estimator has nice properties; see below
- Gauss-Markov assumptions are very strong and often not satisfied
- Relaxations of the Gauss-Markov assumptions and consequences of such relaxations are important topics

Properties of OLS Estimators

1. The OLS estimator b is unbiased: $E\{b\} = \beta$
   Needs assumptions (A1) and (A2)

2. The variance of the OLS estimator b is given by

   $$V\{b\} = \sigma^2 (\Sigma_i x_i x_i')^{-1}$$

   Needs assumptions (A1), (A2), (A3) and (A4)

3. The OLS estimator b is a BLUE (best linear unbiased estimator) for β
   Needs assumptions (A1), (A2), (A3), and (A4) and requires linearity in parameters

Standard Errors of OLS Estimators

Variance of the OLS estimators:

$$V\{b\} = \sigma^2 (\Sigma_i x_i x_i')^{-1}$$

Standard error of OLS estimate b_k: The square root of the k^th^ diagonal element of V{b}

Estimator V{b} is proportional to the variance σ^2 of the error terms

Estimator for σ^2: sampling variance s^2 of the residuals e_i

$$s^2 = (N - K)^{-1} \Sigma_i e_i^2$$

Under assumptions (A1)-(A4), s^2 is unbiased for σ^2

Estimated variance (covariance matrix) of b:

$$s^2 (\Sigma_i x_i x_i')^{-1}$$

Normality of Error Terms

A5	ε_i normally distributed for all i

Together with assumptions (A1), (A3), and (A4), (A5) implies

$\varepsilon_i \sim \text{NID}(0,\sigma^2)$ for all i

i.e., all ε_i are

- independent drawings
- from a *normal* distribution
- with mean 0
- and variance σ^2

Error terms are “normally and independently distributed”

Properties of OLS Estimators

1. The OLS estimator b is unbiased: $E\{b\} = \beta$
2. The variance of the OLS estimator is given by

   $V\{b\} = \sigma^2(\Sigma_i x_i x_i')^{-1}$

3. The OLS estimator b is a BLUE (best linear unbiased estimator) for β

4. The OLS estimator b is normally distributed with mean β and covariance matrix $V\{b\} = \sigma^2(\Sigma_i x_i x_i')^{-1}$

   Needs assumptions (A2) + (A5)

Example: Individual Wages

$$wage_i = \beta_1 + \beta_2\, male_i + \varepsilon_i$$

What do the assumptions mean?

(A1): $\beta_1 + \beta_2\, male_i$ contains the whole systematic part of the model; no regressors besides gender relevant?

(A2): x_i independent of ε_i for all i: knowledge of a person's gender provides no information about further variables which affect the person's wage; is that realistic?

(A3) $V\{\varepsilon_i\} = \sigma^2$ for all i: variance of error terms (and of wages) is the same for males and females; is that realistic?

(A4) $Cov\{\varepsilon_i, \varepsilon_j\} = 0$, $i \neq j$: implied by random sampling

(A5) Normality of ε_i : is that realistic? (Would allow, e.g., for negative wages)

Example: Individ. Wages, cont'd

OLS estimated wage equation (Table 2.1, Verbeek)

Dependent variable: *wage*		
Variable	Estimate	Standard error
constant	5.1469	0.0812
male	1.1661	0.1122
$s = 3.2174$	$R^2 = 0.0317$	$F = 107.93$

$b_1 = 5.147$, se(b_1) = 0.081; $b_2 = 1.166$, se(b_2) = 0.112

95% confidence interval for β_1 : $4.988 \leq \beta_1 \leq 5.306$

Goodness-of-fit

The quality of the linear approximation offered by the model $y_i = x_i'\beta + \varepsilon_i$ can be measured by R^2

- R^2 is the proportion of the variance in y that can be explained by the linear combination of the regressors x_i

$$R^2 = \frac{\hat{V}\{\hat{y}_i\}}{\hat{V}\{y_i\}} = \frac{1/(N-1)\sum_i (\hat{y}_i - \bar{y})^2}{1/(N-1)\sum_i (y_i - \bar{y})^2}$$

- If the model contains an intercept (as usual): $\hat{V}\{y_i\} = \hat{V}\{\hat{y}_i\} + \hat{V}\{e_i\}$

$$R^2 = - \frac{\hat{V}\{e_i\}}{\hat{V}\{y_i\}}$$

- Alternatively, R^2 can be calculated as

$$R^2 = corr^2\{y_i, \hat{y}_i\}$$

Properties of R^2

- $0 \leq R^2 \leq 1$, if the model contains an intercept
- Comparisons of R^2 for two models makes no sense if *y* is different
- R^2 cannot decrease if a variable is added
- *a*djusted R^2: compensated for added regressor, penalty for increasing *K*

$$\bar{R}^2 = 1 - \frac{1/(N-K)\sum_i e_i^2}{1/(N-1)\sum_i (y_i - \bar{y})^2}$$

- Uncentered R^2

$$1 - \Sigma_i e_i^2 / \Sigma_i y_i^2$$

Advanced Econometrics - Lecture 1

- Regression: a descriptive tool
- Economic models
- OLS estimation
- Properties of OLS estimators
- *t*-test and *F*-test
- Asymptotic properties of OLS estimators
- Multicollinearity
- Model specification and tests

Testing of a Regression Coefficient: *t*-Test

For testing a restriction wrt a single regression coefficient β_k:

- Null hypothesis H_0: $\beta_k = q$
- Alternative H_A: $\beta_k > q$ (or $\beta_k < q$ or $\beta_k \neq q$)
- Test statistic: (computed from the sample with known distribution under the null hypothesis)

$$t_k = \frac{b_k - q}{se(b_k)}$$

  t_k follows the t-distribution with N-K degrees of freedom (d.f.)
  - under H_0
  - given the Gauss-Markov assumptions and normality of the error terms ε_i
- Reject H_0, if the ***p*-value** $P\{t_{N-K} > t_k \mid H_0\}$ is small (t_k-value is large)

Example: Individ. Wages, cont'd

OLS estimated wage equation (Table 2.1, Verbeek)

Dependent variable: *wage*		
Variable	Estimate	Standard error
constant	5.1469	0.0812
male	1.1661	0.1122
$s = 3.2174$	$R^2 = 0.0317$	$F = 107.93$

Test of null hypothesis H_0: $\beta_2 = 0$ (no gender effect on wages) against H_A: $\beta_2 > 0$

$$t_2 = b_2/se(b_2) = 1.1661/0.1122 = 10.38$$

Under H_0, t follows the t-distribution with 3294-2 = 3292 d.f.

p-value = $P\{t_{3292} > 10.38 \mid H_0\}$ = 3.7E-25: reject H_0!

Example: Individ. Wages, cont'd

OLS estimated wage equation: Output from GRETL

Modell 1: KQ, benutze die Beobachtungen 1-3294
Abhängige Variable: WAGE

	Koeffizient	*Std. Fehler*	*t-Quotient*	*P-Wert*
const	5,14692	0,0812248	63,3664	<0,00001 ***
MALE	1,1661	0,112242	10,3891	<0,00001 ***

Mittel d. abh. Var.	5,757585	Stdabw. d. abh. Var.	3,269186
Summe d. quad. Res.	34076,92	Stdfehler d. Regress.	3,217364
R-Quadrat	0,031746	Korrigiertes R-Quadrat	0,031452
F(1, 3292)	107,9338	P-Wert(F)	6,71e-25
Log-Likelihood	-8522,228	Akaike-Kriterium	17048,46
Schwarz-Kriterium	17060,66	Hannan-Quinn-Kriterium	17052,82

p-value for t_{MALE}-test: < 0,00001

„gender has a significant effect on wages p.h“

OLS Estimators: Asymptotic Distribution

If the Gauss-Markov (A1) - (A4) assumptions hold but not the normality assumption (A5):

t-statistic

$$t_k = \frac{b_k - \gamma}{se(b_k)}$$

- follows asymptotically ($N \to \infty$) the standard normal distribution

In many situations, the unknown exact properties are substituted by asymptotic results (asymptotic theory)

The *t*-statistic

- follows approximately the *t*-distribution with N-K d.f.
- follows approximately the standard normal distribution N(0,1)

The approximation error decreases with increasing sample size N

Testing Several Regression Coefficients

For testing a restriction wrt more than one, say J with $1<J<K$, regression coefficient:

- Null hypothesis H_0: $\beta_k = 0$, $K-J+1 \leq k \leq K$
- Alternative H_A: at least one of these $\beta_k \neq 0$
- F-statistic: (computed from the sample, with known distribution under the null hypothesis; R_0^2 (R_1^2): R^2 for (un)restricted model)

$$F = \frac{(R_1^2 - R_0^2)/J}{(1 - R_1^2)/(N - K)}$$

  F follows the F-distribution with J and N-K d.f.
  - under H_0
  - given the Gauss-Markov assumptions and normality of the ε_i
- Reject H_0, if the p-value $P\{F_{J,N-K} > F \mid H_0\}$ is small (F-value is large)

Example: Individ. Wages, cont’d

A more general model is

$$wage_i = \beta_1 + \beta_2\, male_i + \beta_3\, school_i + \beta_4\, exper_i + \varepsilon_i$$

β_2 measures the difference in expected wage between a male and a female, given the other regressors fixed, i.e., with the same schooling and experience: ceteris paribus condition

Have *school* <u>and</u> *exper* an explanatory power?

Test of null hypothesis H_0: $\beta_3 = \beta_4 = 0$ against H_A: H_0 not true

- $R_0^2 = 0.0317$ (see p.31)
- $R_1^2 = 0.1326$ (see p.36)

$$F = \frac{(0.1326 - 0.0317)/2}{(1 - 0.1326)/(3294 - 4)} = 191.35$$

- p-value = $P\{F_{2,3290} > 191.35 \mid H_0\} = 2.43\text{E-}79$

Example: Individ. Wages, cont'd

OLS estimated wage equation (Table 2.2, Verbeek)

Table 2.2 OLS results wage equation

Dependent variable: *wage*

Variable	Estimate	Standard error	t-ratio
constant	−3.3800	0.4650	−7.2692
male	1.3444	0.1077	12.4853
school	0.6388	0.0328	19.4780
exper	0.1248	0.0238	5.2530

$s = 3.0462 \quad R^2 = 0.1326 \quad \bar{R}^2 = 0.1318 \quad F = 167.63$

Testing Several Regression Coefficients, cont'd

Test again

- H_0: $\beta_k = 0$, K-J+1 ≤ k ≤ K
- H_A: at least one of these $\beta_k \neq 0$

The test statistic F can alternatively be calculated as

$$F = \frac{(S_0 - S_1)/J}{S_1/(N-K)}$$

S_0 (S_1): sum of squared residuals for the (un)restricted model

Example: Individ. Wages, cont'd

A more general model is

$$wage_i = \beta_1 + \beta_2\ male_i + \beta_3\ school_i + \beta_4\ exper_i + \varepsilon_i$$

β_2 measures the difference in expected wage between a male and a female, given the other regressors fixed, i.e., with the same schooling and experience: **ceteris paribus condition**

Have *school* and *exper* an explanatory power?

Test of null hypothesis H_0: $\beta_3 = \beta_4 = 0$ against H_A: H_0 not true

- $S_0 = 34076.92$ (see p.32)
- $S_1 = 30527.87$

$$F = [(34076.92 - 30527.87)/2]/[30527.87/(3294-4)] = 191.24$$

Does <u>any</u> regressor contribute to explanation? Overall *F*-test (see Table 2.2 or GRETL-output): $F = 167.63$, *p*-value: 4.0E-101

The General Case

Test of H_0: $R\beta = q$

$R\beta = q$: J linear restrictions on the coefficients (R: JxK matrix, q: K-vector)

Example: $$R = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 0 & 1 & -1 & 1 \end{pmatrix}, \quad q = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Wald test: $\xi = (Rb-q)'[RV\{b\}R']^{-1}(Rb-q)$ follows under H_0 approximately the Chi-squared distribution with J d.f.

$F = \xi / J$ is algebraically identical to the F-test with

$$F = \frac{(S_0 - S_1)/J}{S_1/(N-K)}$$

p-value, Size, and Power

Type I error: the null hypothesis is rejected, while it is actually true

- ***p*-value**: the probability to commit the type I error
- In experimental situations, the probability of committing the type I error can be chosen before applying the test; the probability of committing the type I error is denoted the **size α of the test**
- In model-building situations, not a decision but learning from data is intended; multiple testing is quite usual; use of *p*-value is more appropriate

Type II error: the null hypothesis is not rejected, while it is actually wrong

- The probability to decide in favor of the true alternative, i.e., not making a type II error, is called the **power** of the test; depends of true parameter values

p-value, Size, and Power, cont'd

- The smaller the size of the test, the larger is its power (for a given sample size)
- The more H_A deviates from H_0, the larger is the power of a test of a given size (given the sample size)
- The larger the sample size, the larger is the power of a test of a given size

Attention! Significance vs relevance

Advanced Econometrics - Lecture 1

- Regression: a descriptive tool
- Economic models
- OLS estimation
- Properties of OLS estimators
- *t*-test and *F*-test
- Asymptotic properties of OLS estimators
- Multicollinearity
- Model specification and tests

OLS Estimators: Asymptotic Properties

Gauss-Markov assumptions plus the normality assumptions are in many situations very restrictive

An alternative are properties derived from asymptotic theory

- Asymptotic results hopefully are sufficiently precise approximations for large (but finite) N
- Typically, Monte Carlo simulations are used to assess the quality of asymptotic results

Asymptotic theory: deals with the case where the sample size N goes to infinity: $N \to \infty$

OLS Estimators: Consistency

Consistency of the OLS estimators *b*:

- For $N \to \infty$, the probability that *b* differs from β by a certain amount goes to 0
- The distribution of *b* collapses in β

The OLS estimators *b* are consistent,

$\text{plim}_{N \to \infty}\, b = \beta$,

if (A2) from the Gauss-Markov assumptions and the assumption (A6) is fulfilled:

A6	$1/N\ \Sigma^{N}_{i=1}\, x_i\, x_i'$ converges with growing *N* to a finite, nonsingular matrix Σ_{xx}

OLS Estimators: Consistency, cont'd

Consistency of the OLS estimators can also be shown to hold under weaker assumptions:

The OLS estimators b are consistent,

$\text{plim}_{N \to \infty}\, b = \beta$,

if the assumptions (A7) and (A6) are fulfilled

A7	The error terms have zero mean and are uncorrelated with each of the regressors: $E\{x_i\, \varepsilon_i\} = 0$

Attention:

- (A7) does not imply (A2)
- The conditions for consistency are weaker than that for unbiasedness

OLS Estimators: Consistency, cont'd

The estimator s^2 for the error term variance σ^2 is consistent,

$$\text{plim}_{N \to \infty} s^2 = \sigma^2,$$

if the assumptions (A3), (A6), and (A7) are fulfilled

OLS Estimators: Asymptotic Normality

Under the Gauss-Markov assumptions (A1)-(A4) and assumption (A6), the OLS estimators *b* follow approximately the normal distribution

$$N\left(\beta, s^2 \left(\sum_i x_i x_i'\right)^{-1}\right)$$

The approximate distribution does not make use of assumption (A5), i.e., the normality of the error terms!

Tests of hypotheses on coefficients β_k,

- *t*-test
- *F*-test

can be perfomed making use of the approximate normal distribution

Advanced Econometrics - Lecture 1

- Regression: a descriptive tool
- Economic models
- OLS estimation
- Properties of OLS estimators
- *t*-test and *F*-test
- Asymptotic properties of OLS estimators
- Multicollinearity
- Model specification and tests

Multicollinearity

OLS estimators $b = (X'X)^{-1}X'y$ for regression coefficients β require that the KxK matrix

$$X'X \text{ or } \Sigma_i\, x_i\, x_i'$$

can be inverted

In real situations, regressors may be correlated, such as

- experience and schooling (measured in years)
- age and experience
- inflation rate and nominal interest rate
- common trends of economic time series, e.g., in lag structures

Multicollinearity: between the explanatory variables exists

- an exact linear relationship
- an approximate linear relationship

Multicollinearity: Consequences

Exact linear relationship between regressors ("exact multicollinearity"):

- Example: Wage equation
  - Regressors *male* <u>and</u> *female* in addition to *intercept*
  - Regressor *exper* defined as *exper* = *age* - *school* - 6
- $\Sigma_i x_i x_i'$ is not invertible
- Econometric software reports ill-defined matrix $\Sigma_i x_i x_i'$
- GRETL drops regressor

Approximate linear relationship between regressors:

- When correlations are high: hard to identify the *individual* impact of each of the regressors
- Inflated variances: if x_k can be approximated by the other regressors, variance of b_k is inflated; reduced power of *t*-test

Variance Inflation Factor

Variance of b_k

$$V\{b_k\} = \frac{\sigma^2}{1-R_k^2}\frac{1}{N}\left[\frac{1}{N}\sum_{i}^{N}(x_{ik}-\bar{x}_k)^2\right]^{-1}$$

R_k^2: R^2 of the regression of x_k on all other regressors

- If x_k can be approximated by the other regressors, R_k^2 is close to 1, the variance inflated

Variance inflation factor: $\mathrm{VIF}(b_k) = (1 - R_k^2)^{-1}$

Large values for some or all VIFs indicate multicollinearity

Attention! Large values for VIF can also have other causes

- Small value of variance of X_k
- Small number N of observations

Multicollinearity: Indicators

Large values for some or all variance inflation factors VIF(b_k) are an indicator for multicollinearity

Other indicators:

- At least one of the R_k^2, k = 1, …, K, has a large value
- Large values of standard errors se(b_k) (low t-statistics), but reasonable or good R^2 and F-statistics
- Effect of adding a regressor on standard errors se(b_k) of estimates b_k of regressors already in the model: increasing values of se(b_k) indicate multicollinearity

Advanced Econometrics - Lecture 1

- Regression: a descriptive tool
- Economic models
- OLS estimation
- Properties of OLS estimators
- *t*-test and *F*-test
- Asymptotic properties of OLS estimators
- Multicollinearity
- Model specification and tests

Selection of Regressors

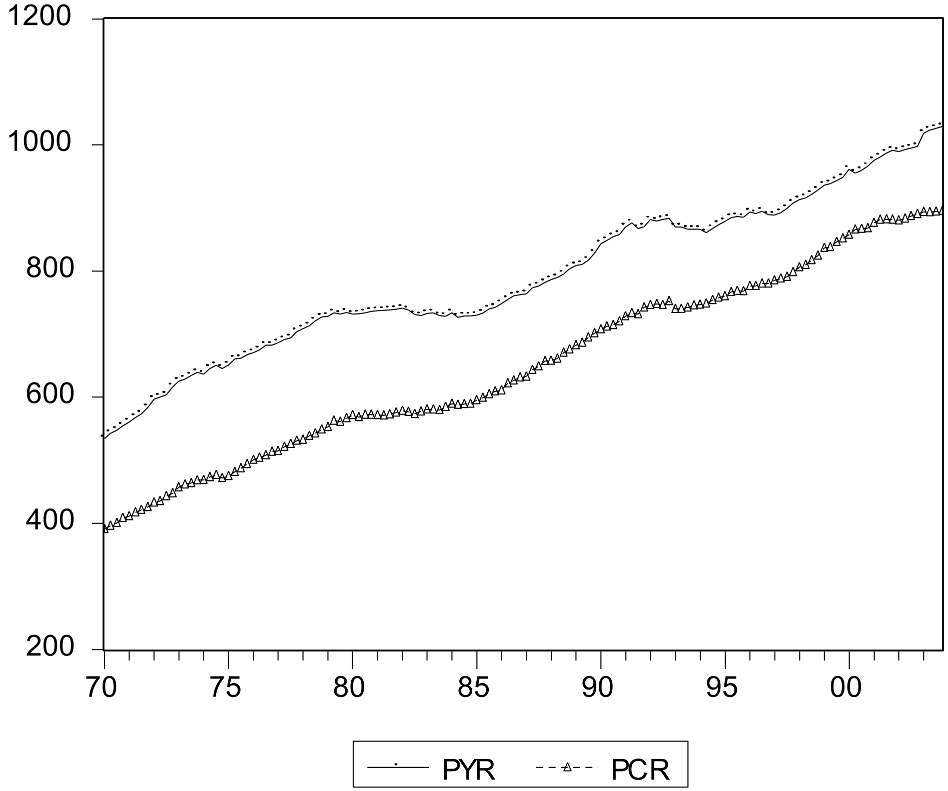
Specification errors:

- Omission of a relevant variable
- Inclusion of a irrelevant variable

Questions:

- What are the consequences?
- How to avoid specification errors?
- How to detect a committed specification error?

Example: Income and Consumption



PCR: Private Consumption, real, in bn. EUROs

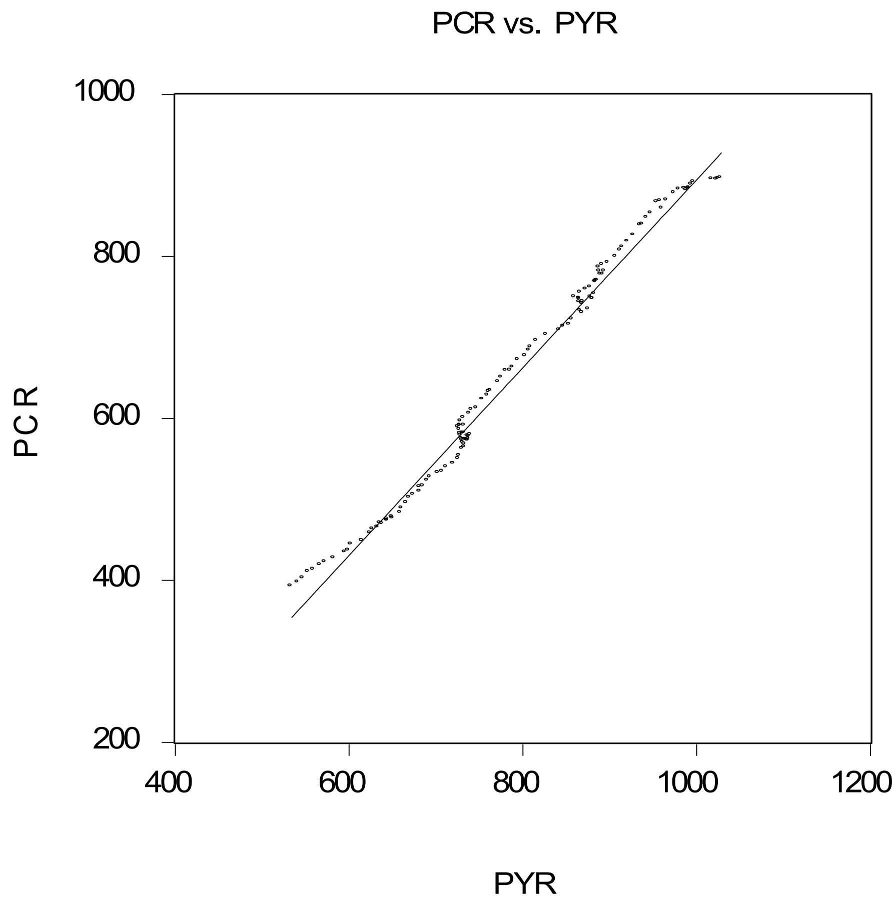
PYR: Household's Disposable Income, real, in bn. EUROs

1970:1-2003:4

Basis: 1995

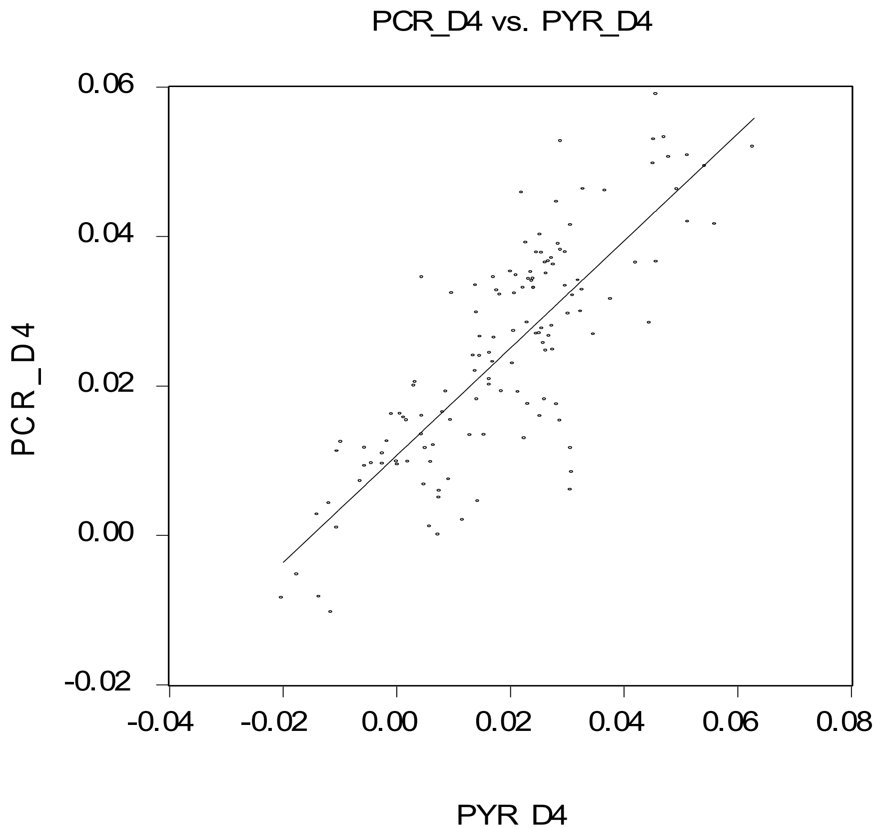
Source: AWM-Database

Income and Consumption



PCR: Private Consumption, real, in bn. EUROs
PYR: Household's Disposable Income, real, in bn. EUROs
1970:1-2003:4
Basis: 1995
Source: AWM-Database

Income and Consumption: Growth Rates



PCR_D4: Private Consumption, real, growth rate
PYR_D4: Household's Disposable Income, real, growth rate
1970:1-2003:4
Basis: 1995
Source: AWM-Database

Consumption Function

C: Private Consumption, real, growth rate (PCR_D4)

Y: Household's Disposable Income, real, growth rate (PYR_D4)

T: Trend ($T_i = i/1000$)

$$\hat{C} = 0.011 + 0.761\,Y, \quad \bar{R}^2 = 0.717$$

Consumption function with trend T_i = i/1000:

$$\hat{C} = 0.016 + 0.708\,Y - 0.068\,T, \quad \bar{R}^2 = 0.741$$

Consumption Function, cont'd

OLS estimated consumption function: Output from GRETL

Abhängige Variable: PCR_D4

	Koeffizient	Std.-fehler	t-Quotient	P-Wert
const	0,0162489	0,00187868	8,649	1,76e-014 ***
PYR_D4	0,707963	0,0424086	16,69	4,94e-034 ***
T	-0,0682847	0,0188182	-3,629	0,0004 ***

Mittel d. abh. Var.	0,024911	Stdabw. d. abh. Var.	0,015222
Summe d. quad. Res.	0,007726	Stdfehler d. Regress.	0,007739
R-Quadrat	0,745445	Korrigiertes R-Quadrat	0,741498
F(2, 129)	188,8830	P-Wert(F)	4,71e-39
Log-Likelihood	455,9302	Akaike-Kriterium	-905,8603
Schwarz-Kriterium	-897,2119	Hannan-Quinn-Kriterium	-902,3460
rho	0,701126	Durbin-Watson-Stat	0,601668

Selection of Regressors

Specification errors:

- Omission of a relevant variable
- Inclusion of a irrelevant variable

Questions:

- What are the consequences?
- How to avoid specification errors?
- How to detect a committed specification error?

Misspecification: Omitted Regressor

Two models:

$y_i = x_i'\beta + z_i'\gamma + \varepsilon_i$ (A)

$y_i = x_i'\beta + v_i$ (B)

OLS estimates b_B of β from (B) can be written with y_i from (A):

$$b_B = \beta + \left(\sum_i x_i x_i'\right)^{-1} \sum_i x_i z_i'\gamma + \left(\sum_i x_i x_i'\right)^{-1} \sum_i x_i \varepsilon_i$$

If (A) is the true model but (B) is specified, i.e., relevant regressors z_i are omitted, b_B is biased by

$$E\left\{\left(\sum_i x_i x_i'\right)^{-1} \sum_i x_i z_i'\gamma\right\}$$

Omitted variable bias

No bias if (a) γ = 0 or if (b) variables in x_i and z_i are orthogonal

Misspecification: Irrelevant Regressor

Two models:

$y_i = x_i'\beta + z_i'\gamma + \varepsilon_i$ (A)

$y_i = x_i'\beta + v_i$ (B)

If (B) is the true model but (A) is specified, i.e., the model contains irrelevant regressors z_i

The OLS estimates b_A

- are unbiased
- have a higher variance than the OLS estimate b_B obtained from fitting model (B)

Specification Search

General-to-specific modeling:

1. List all potential regressors
2. Specify the most general model: it includes all potential regressors
3. Test iteratively which variables have to be dropped
4. Stop if no more variables have to be dropped

The procedure is also known as the LSE (London School of Economics) method

Specification Search, cont'd

Some remarks

- Alternatively, one can start with a small model and add variables as long as they turn out to contribute to explaining Y
- Stepwise regression
- Adding and deleting can be based on
  - t-statistic, F-statistic
  - Adjusted R^2
  - Akaike's Information Criterion AIC, Schwarz's Bayesian Information Criterion BIC
- The corresponding probabilities for type I and type II errors can hardly be assessed

Specification search can be subsumed under **data mining**

Comparison of Models

Nested models [cf. p.58: model (B) is nested in model (A)]

- Do the *J* added regressors contribute to explaining *Y*
- *F*-test (*t*-test when *J* = 1) for testing H_0: coefficients of added regressors are zero

$$F = \frac{(R_1^2 - R_0^2)/J}{(1 - R_1^2)/(N - K)}$$

R_0^2 and R_1^2 are the R^2 of the models without and with the *J* additional regressors, respectively

- Comparison of adjusted R^2: adj R_1^2 > adj R_0^2 equivalent to $F > 1$
- Information Criteria: penalty for increasing number of regressors (cf. adjusted R^2), e.g., Schwarz's Bayesian Information Criterion

$$BIC = \log \tfrac{1}{N} \sum_{\iota} {}_i^2 + \tfrac{K}{N} \log N$$

Comparison of Models, cont'd

Non-nested alternative models: A: $y_i = x_i'\beta + \varepsilon_i$, B: $y_i = z_i'\gamma + v_i$

- Non-nested or encompassing F-test: compares by F-tests artificially nested models

  $y_i = x_i'\beta + z_{2i}'\delta_B + \varepsilon_i$ with z_{2i} not element of x_i: test of $\delta_B = 0$

- J-test: applies an F-test to a combined model

  $y_i = (1 - \delta)\, x_i'\beta + \delta\, z_{2i}'\gamma + u_i$

Choice between linear and loglinear functional form

- PE-test

PE-Test

- Estimate both models
  - A: $y_i = x_i'\beta + \varepsilon_i$
  - B: $\log y_i = x_i'\beta + v_i$

  and calculate the fitted values $\hat{y}$ (from model A) and $\ddot{y}$ (from B)
- Test $\delta_{LIN} = 0$ in

  $$y_i = x_i'\beta + \delta_{LIN} (\log \hat{y}_i - \log \ddot{y}_i) + u_i$$

  not rejecting $\delta_{LIN} = 0$ favors the linear model
- Test $\delta_{LOG} = 0$ in

  $$\log y_i = x_i'\beta + \delta_{LOG} (\hat{y}_i - \exp\{\log \ddot{y}_i\}) + u_i$$

  not rejecting $\delta_{LOG} = 0$ favors the linear model
- Rejection both null hypotheses: find a more adequate model

Testing the Functional Form

Misspecification of $y_i = x_i'\beta + \varepsilon_i$: violation of linearity in x_i

- $E\{y_i|x_i\} = g(x_i, \beta)$, e.g.,
  - $g(x_i, \beta) = \beta_1 + \beta_2\, x_i^{\beta 3}$
  - $g(x_i, \beta) = \beta_1\, x_{i1}^{\beta 2}\, x_{i2}^{\beta 3}$
- Linear model $x_i'\beta$ does not explain well Y
- RESET (Regression Equation Specification Error Test) test (Ramsey)
  - Alternative model: linear model extended by adding $\hat{y}_i^2$, $\hat{y}_i^3$, ... with $\hat{y}_i$: fitted values from the linear model
  - Uses F-test to decide whether powers of fitted values contribute as additional regressors to explaining Y
  - Power Q of fitted values: typical choice is $Q = 2$ or $Q = 3$

Exercise

In Exercise 2.2 of Verbeek, the sample given in data set “wages” is used to answer the question whether women are systematically underpaid compared with men. Table 2.8, p.48, gives the output of a regression analysis, the model for the log hourly wages being explained besides *male* by *age* and *educ*. Use in this exercise the whole dataset (data file *WAGES1*) and the definition *age* = *school* + *exper* + *6*.

1. Repeat the analysis for the model (model 1) where the log hourly wages are explained by *male* and *age*.
2. Repeat the analysis (model 2) after adding to model 1 four dummy variables for the educational levels 2 through 5 instead of the variable *educ.*

Exercise, cont'd

3. Use an *F*-test, adjusted R^2, and the BIC to decide whether model 1 or that model 2 is preferable.
4. Use the PE-test (see Verbeek, p. 64) to decide whether the Verbeek's model in Table 2.8 (where levels of hourly wages are explained) or the model 1 extended by the variable *educ* is to be preferred.