Econometrics - Lecture 2

Introduction to Linear Regression – Part 2

Contents

Goodness-of-fit R^2

The quality of the model $y_i = x_i'\beta + \varepsilon_i$, $i = 1, \ldots, N$, with K regressors can be measured by R^2, the goodness-of-fit (GoF) statistic

- R^2 is the portion of the variance in Y that can be explained by the linear regression with regressors X_k, $k=1,\ldots,K$

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

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

$$R^2 = 1 - \frac{\hat{V}\{e_i\}}{\hat{V}\{y_i\}}$$

with $\hat{V}\{e_i\}$ = $(\Sigma_i e_i^2)/(N-1)$

- Alternatively, R^2 can be calculated as

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

Properties of R^2

R^2 is the portion of the variance in Y that can be explained by the linear regression; $100R^2$ is measured in percent

- $0 \leq R^2 \leq 1$, if the model contains an intercept
- $R^2 = 1$: all residuals are zero
- $R^2 = 0$: for all regressors, $b_k = 0$, $k = 2, \ldots, K$; the model explains nothing
- R^2 cannot decrease if a variable is added
- Comparisons of R^2 for two models makes no sense if the explained variables are different

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
$s = 3.2174$	$R^2 = 0.0317$	$F = 107.93$

only 3.17% of the variation of individual wages p.h. is due to the gender

Individual Wages, cont'd

Wage equation with three regressors (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

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

R^2 increased due to adding *school* and *exper*

Other GoF Measures

- Uncentered R^2: For the case of no intercept; the Uncentered R^2 cannot become negative

  Uncentered $R^2 = 1 - \Sigma_i e_i^2 / \Sigma_i y_i^2$
- adj R^2 (adjusted R^2): For comparing models; compensated for added regressor, penalty for increasing K

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

  for a given model, *adj* R^2 is smaller than R^2
- For other than OLS estimated models

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

  it coincides with R^2 for OLS estimated models

Contents

- Goodness-of-Fit
- Hypothesis Testing
- Testing Linear Restrictions
- Asymptotic Properties of the OLS Estimator
- Multicollinearity
- Prediction

Individual Wages

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: mean wage p.h. for females: 5.15\$, with std.error of 0.08\$

$b_2 = 1.166$, se(b_2) = 0.112

OLS Estimator: Distributional Properties

Under the assumptions (A1) to (A5):

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

  $$b \sim N(\beta, \sigma^2(X'X)^{-1}), \quad b_k \sim N(\beta_k, \sigma^2 c_{kk}), \; k=1,\ldots,K$$

  with c_{kk} the k-th diagonal element of $(X'X)^{-1}$
- The statistic

  $$z = \frac{b_k - \beta_k}{se(b_k)} = \frac{b_k - \beta_k}{\sigma\sqrt{c_{kk}}}$$

  follows the standard normal distribution N(0,1)
- The statistic

  $$t_k = \frac{b_k - \beta_k}{s\sqrt{c_{kk}}}$$

  follows the t-distribution with N-K degrees of freedom (*df*)

Testing a Regression Coefficient: *t*-Test

For testing a restriction on the (single) regression coefficient β_k:

- Null hypothesis H_0: $\beta_k = q$ (most interesting case: $q = 0$)
- Alternative H_A: $\beta_k > q$
- Test statistic: (computed from the sample with known distribution under the null hypothesis)

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

- t_k is a realization of the random variable t_{N-K}, which follows the *t*-distribution with *N-K* degrees of freedom ($df = N-K$)
  - under H_0 and
  - given the Gauss-Markov assumptions and normality of the errors
- Reject H_0, if the *p*-value $P\{t_{N-K} > t_k \mid H_0\}$ is small (t_k-value is large)

Individual 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, equal wages for males and females) against H_A: $\beta_2 > 0$

$$t_2 = b_2/\text{se}(b_2) = 1.1661/0.1122 = 10.38$$

Under H_0, T follows the t-distribution with df = 3294-2 = 3292

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

Individual Wages, cont'd

OLS estimated wage equation: Output from GRETL

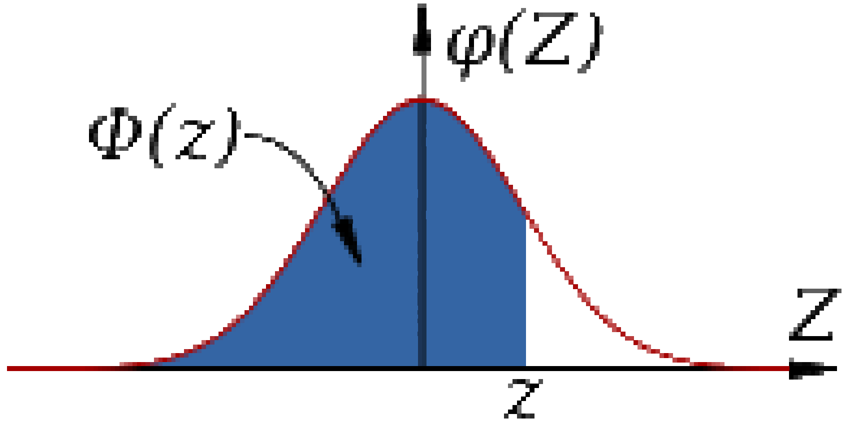
Model 1: OLS, using observations 1-3294
Dependent variable: WAGE

	coefficient	*std. error*	*t-ratio*	*p-value*
const	5,14692	0,0812248	63,3664	<0,00001 ***
MALE	1,1661	0,112242	10,3891	<0,00001 ***

Mean dependent var	5,757585	S.D. dependent var	3,269186
Sum squared resid	34076,92	S.E. of regression	3,217364
R- squared	0,031746	Adjusted R- squared	0,031452
F(1, 3292)	107,9338	P-value(F)	6,71e-25
Log-likelihood	-8522,228	Akaike criterion	17048,46
Schwarz criterion	17060,66	Hannan-Quinn	17052,82

p-value for t_{MALE}-test: < 0.00001
„gender has a significant effect on wages, males earn more"

Normal and *t*-Distribution



Standard normal distribution: $Z \sim N(0,1)$

- Distribution function $\Phi(z) = P\{Z \leq z\}$

t-distribution: $T_{df} \sim t(df)$

- Distribution function $F(t) = P\{T_{df} \leq t\}$
- *p*-value: $P\{T_{N-K} > t_k \mid H_0\} = 1 - F_{H0}(t_k)$

For growing *df*, the *t*-distribution approaches the standard normal distribution, T_{df} follows asymptotically ($N \to \infty$) the N(0,1)-distribution

- 0.975-percentiles $t_{df,0.975}$ of the $t(df)$-distribution

df	5	10	20	30	50	100	200	∞
$t_{df,0.025}$	2.571	2.228	2.085	2.042	2.009	1.984	1.972	1.96

- 0.975-percentile of the standard normal distribution: $z_{0.975} = 1.96$

OLS Estimators: Asymptotic Distribution

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

t-statistic

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

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

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

The *t*-statistic

- follows 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

Two-sided *t*-Test

For testing a restriction wrt a single regression coefficient β_k:

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

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

  follows the *t*-distribution with N-K d.f.
- Reject H_0, if the *p*-value

$$P\{|T_{N-K}| > |t_k| \mid H_0\}$$

  is small ($|t_k|$-value is large)

Individual 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, equal wages for males and females) against H_A: $\beta_2 \neq 0$

$$t_2 = b_2/\text{se}(b_2) = 1.1661/0.1122 = 10.38$$

Under H_0, T follows the t-distribution with df = 3294-2 = 3292

p-value = $P\{T_{3292} < -10.38 \text{ or } T_{3292} > 10.38 \mid H_0\}$ = 7.4E-25: reject H_0!

Significance Tests

For testing a restriction wrt a single regression coefficient β_k:

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

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

- Determine the critical value $t_{N-K,1-\alpha/2}$ for the significance level α from

$$P\{|T_k| > t_{N-K,1-\alpha/2} \mid H_0\} = \alpha$$

- Reject H_0, if $|t_k| > t_{N-K,1-\alpha/2}$
- Typically, the value 0.05 is taken for α

Significance Tests, cont'd

One-sided test :

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

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

- Determine the critical value $t_{N-K,\alpha}$ for the significance level α from

$$P\{T_k > t_{N-K,\alpha} \mid H_0\} = \alpha$$

- Reject H_0, if $t_k > t_{N-K,\alpha}$ ($t_k < -t_{N-K,\alpha}$)

Confidence Interval for β_k

Range of values (b_{kl}, b_{ku}) for which the null hypothesis on β_k is not rejected

$$b_{kl} = b_k - t_{N-K,1-\alpha/2}\,\mathrm{se}(b_k) < \beta_k < b_k + t_{N-K,1-\alpha/2}\,\mathrm{se}(b_k) = b_{ku}$$

- Refers to the significance level α of the test
- For large values of *df* and $\alpha = 0.05$ ($1.96 \approx 2$)

$$\boxed{b_k - 2\,\mathrm{se}(b_k) < \beta_k < b_k + 2\,\mathrm{se}(b_k)}$$

- Confidence level: $\gamma = 1 - \alpha$; typically $\gamma = 0.95$

Interpretation:

- A range of values for the true β_k that are not unlikely (contain the true value with probability $100\gamma\%$), given the data (?)
- A range of values for the true β_k such that $100\gamma\%$ of all intervals constructed in that way contain the true β_k

Individual 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
$s = 3.2174$	$R^2 = 0.0317$	$F = 107.93$

The confidence interval for the gender wage difference (in USD p.h.)

- confidence level $\gamma = 0.95$

  $1.1661 - 1.96*0.1122 < \beta_2 < 1.1661 + 1.96*0.1122$

  $0.946 < \beta_2 < 1.386$ (or **0.94** $< \beta_2 < 1.39$)
- $\gamma = 0.99$: $0.877 < \beta_2 < 1.455$

Contents

Testing a Linear Restriction on Regression Coefficients

Linear restriction $r'\beta = q$

- Null hypothesis H_0: $r'\beta = q$
- Alternative H_A: $r'\beta > q$
- Test statistic

$$t = \frac{r'b - q}{se(r'b)}$$

  $se(r'b)$ is the square root of $V\{r'b\} = r'V\{b\}r$

- Under H_0 and (A1)-(A5), t follows the t-distribution with $df = N\text{-}K$

GRETL: The option Linear restrictions from Tests on the output window of the Model statement Ordinary Least Squares allows to test linear restrictions on the regression coefficients

Testing Several Regression Coefficients: *F*-test

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

- Null hypothesis H_0: $\beta_k = 0$, $K\text{-}J+1 \le k \le K$
- Alternative H_A: for at least one k, $K\text{-}J+1 \le k \le K$, $\beta_k \ne 0$
- F-statistic: (computed from the sample, with known distribution under H_0; R_0^2: R^2 for restricted model; R_1^2: R^2 for unrestricted model)

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

  F follows the F-distribution with J and $N\text{-}K$ d.f.

  - under H_0 and given the Gauss-Markov assumptions (A1)-(A4) and normality of the ε_i (A5)
- Reject H_0, if the p-value $P\{F_{J,N-K} > F \mid H_0\}$ is small (F-value is large)
- The F-test with $J = K$-1 is a standard test in GRETL

Individual 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 wages p.h. between males and females, given the other regressors fixed, i.e., with the same schooling and experience: ceteris paribus condition

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

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

- $R_0^2 = 0.0317$
- $R_1^2 = 0.1326$

$$F = \frac{(0.1326 - 0.0317)/2}{(1 - 0.1326)/(3294 - 4)} = 191.24$$

- p-value = $P\{F_{2,3290} > 191.24 \mid H_0\} = 2.68E\text{-}79$

Individual 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

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

Alternatives for Testing Several Regression Coefficients

Test again

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

1. 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
- F follows under H_0 and (A1)-(A5) the $F(J,N-K)$-distribution

2. If σ^2 is known, the test can be based on

$$F = (S_0 - S_1)/\sigma^2$$

under H_0 and (A1)-(A5): Chi-squared distributed with J d.f.

- For large N, s^2 is very close to σ^2; test with F approximates F-test

Individual Wages, cont'd

A more general model is

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

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
- $S_0 = 34076.92$, $S_1 = 30527.87$
- $s = 3.046143$

$$F_{(1)} = [(34076.92 - 30527.87)/2]/[30527.87/(3294\text{-}4)] = 191.24$$

$$F_{(2)} = [(34076.92 - 30527.87)/2]/3.046143 = 191.24$$

Does <u>any</u> regressor contribute to explanation?

- Overall *F*-test for H_0: $\beta_2 = \ldots = \beta_4 = 0$ against H_A: H_0 not true (see Table 2.2 or GRETL-output): J=3

$$F = 167.63,\ p\text{-value: } 4.0\text{E-}101$$

The General Case

Test of H_0: $R\beta = q$

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

Example:

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

Wald test: test statistic

$$\xi = (Rb - q)'[RV\{b\}R']^{-1}(Rb - q)$$

- follows under H_0 for large N approximately the Chi-squared distribution with J d.f.
- Test based on $F = \xi / J$ 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; this probability is the significance level α, also denoted as the **size of the test**
- In model-building situations, not a decision but learning from data is intended; multiple testing is quite usual; the use of *p*-values is more appropriate than using a strict α

Type II error: the null hypothesis is not rejected, while it is actually wrong; the decision is not in favor of the true alternative

- 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 smaller 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

Contents

OLS Estimators: Asymptotic Properties

Gauss-Markov assumptions (A1)-(A4) plus the normality assumption (A5) 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: $N \to \infty$

Chebychev's Inequality

Chebychev's Inequality: Bound for the probability of deviations from its mean

$$P\{|z-\mathrm{E}\{z\}| > r\sigma\} < r^{-2}$$

for all $r>0$; true for any distribution with moments E{z} and σ^2 = V{z}

For OLS estimator b_k:

$$\mathrm{P}\{|\,\mathrm{b_k} - \beta_\mathrm{k}\,| > \delta\} < \frac{\sigma^2 c_{kk}}{\delta^2}$$

for all $\delta>0$; c_{kk}: the k-th diagonal element of $(X'X)^{-1} = (\Sigma_i x_i x_i')^{-1}$

- For growing N: the elements of $\Sigma_i x_i x_i'$ increase, V{b_k} decreases
- Given (A6) [see next slide], for all $\delta>0$

$$\lim_{N\to\infty} P\{|\,b_k - \beta_k\,| > \delta\} = 0$$

b_k converges in probability to β_k for $N \to \infty$; $\mathrm{plim}_{N\to\infty}\, b_k = \beta_k$

Consistency of the OLS-estimator

Simple linear regression

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

Observations: (y_i, x_i), $i = 1, \ldots, N$

OLS estimator

$$b_2 = \left[\sum_{i=1}^{N}(x_i - \bar{x})y_i\right] / \left[\sum_{i=1}^{N}(x_i - \bar{x})^2\right]$$

$$= \beta_2 + \left[N^{-1}\sum_{i=1}^{N}(x_i - \bar{x})\varepsilon_i\right] / \left[N^{-1}\sum_{i=1}^{N}(x_i - \bar{x})^2\right]$$

- $N^{-1}\sum_{i=1}^{N}(x_i - \bar{x})\varepsilon_i$ and $N^{-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$ converge in probability to Cov $\{x, \varepsilon\}$ and V$\{x\}$, respectively
- Due to (A2), Cov $\{x, \varepsilon\} = 0$; with V$\{x\} > 0$ follows

$$\text{plim}_{N \to \infty} b_2 = \beta_2 + \text{Cov}\,\{x, \varepsilon\}/\text{V}\{x\} = \beta_2$$

OLS Estimators: Consistency

If (A2) from the Gauss-Markov assumptions (exogenous x_i, all x_i and ε_i are independent) and the assumption (A6) are fulfilled:

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

b_k converges in probability to β_k for $N \rightarrow \infty$

Consistency of the OLS estimators b:

- For $N \rightarrow \infty$, b converges in probability to β, i.e., the probability that b differs from β by a certain amount goes to zero for $N \rightarrow \infty$
- The distribution of b collapses in β
- $\text{plim}_{N \rightarrow \infty}\ b = \beta$

Needs no assumptions beyond (A2) and (A6)!

OLS Estimators: Consistency, cont'd

Consistency of 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: $E\{x_i\, \varepsilon_i\} = 0$

Follows from

$$b = \beta + \left(\frac{1}{N} \sum_i x_i x_i' \right)^{-1} \frac{1}{N} \sum_i x_i \varepsilon_i$$

and

$$\text{plim}(b - \beta) = \Sigma_{xx}^{-1} E\{x_i\, \varepsilon_i\}$$

Consistency of s^2

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

Consistency: Some Properties

- plim g(b) = g(β)
  - if plim $s^2 = \sigma^2$, then plim $s = \sigma$
- The conditions for consistency are weaker than those for unbiasedness

OLS Estimators: Asymptotic Normality

- Distribution of OLS estimators mostly unknown
- Approximate distribution, based on the asymptotic distribution
- Many estimators in econometrics follow asymptotically the normal distribution
- Asymptotic distribution of the consistent estimator *b*: distribution of

  $N^{1/2}(b - \beta)$ for $N \to \infty$

- Under the Gauss-Markov assumptions (A1)-(A4) and assumption (A6), the OLS estimators *b* fulfill

  $$\sqrt{N}(b-\beta) \to \mathrm{N}\left(0, \sigma^2 \Sigma_{xx}^{-1}\right)$$

  "→" means "is asymptotically distributed as"

OLS Estimators: Approximate Normality

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

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

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 performed by making use of the approximate normal distribution

Assessment of Approximate Normality

Quality of

- approximate normal distribution of OLS estimators
- *p*-values of *t*- and *F*-tests
- power of tests, confidence intervals, etc.

  depends on sample size *N* and factors related to Gauss-Markov assumptions etc.

Monte Carlo studies: simulations that indicate consequences of deviations from ideal situations

Example: $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$; distribution of b_2 under classical assumptions?

- 1) Choose *N*; 2) generate x_i, ε_i, calculate y_i, i=1,…,*N*; 3) estimate b_2
- Repeat steps 1)-3) *R* times: the *R* values of b_2 allow assessment of the distribution of b_2

Contents

Individual Wages: Variable *Age*

Define the variable

$$age_i = 6 + school_i + exper_i$$

For the model

$$wage_i = \beta_1 + \beta_2\, male_i + \beta_3\, age_i + \beta_4\, school_i + \beta_5\, exper_i + \varepsilon_i$$

- the Nx5 design matrix X has rank 4
- it has not full rank 5!
- $X'X$ cannot be inverted

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

- age and experience (measured in years)
- experience and schooling
- 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 (exact collinearity)
- an approximate linear relationship

Multicollinearity: Consequences

Approximate linear relationship between regressors:

- When correlations between regressors are high: difficult 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;
  - Smaller t_k-statistic, reduced power of t-test
- Example: $y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$
  - with sample variances of X_1 and X_2 equal 1 and correlation r_{12},

$$V\{b\} = \frac{\sigma^2}{N} \frac{1}{1-r_{12}^2} \begin{pmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{pmatrix}$$

r_{12}	0,3	0,5	0,7	0,9
$1/(1-r_{12}^2)$	1,10	1,33	1,96	5,26

Exact Collinearity

Exact linear relationship between regressors

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

Remedy:

- Exclude (one of the) regressors
- Example: Wage equation, *male* <u>and</u> *female* in addition to *intercept*
  - Drop regressor *female,* use only regressor *male* in addition to *intercept*
  - Alternatively: use *female* and *intercept*
  - Not good: use of *male* and *female*, no *intercept*

Variance Inflation Factor

Variance of b_k

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

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

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

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

Large values for some or all VIFs indicate multicollinearity

Warning! Large values of the variance of b_k (and reduced power of the *t*-test) can have various causes

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

Other Indicators for Multicollinearity

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-statistic
- 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

Contents

- Goodness-of-Fit
- Hypothesis Testing
- Testing Linear Restrictions
- Asymptotic Properties of the OLS Estimator
- Multicollinearity
- Prediction

The Predictor

Given the relation $y_i = x_i'\beta + \varepsilon_i$

Given estimators b, predictor for the expected value of Y at x_0, i.e.,

$y_0 = x_0'\beta + \varepsilon_0$: $\hat{y}_0 = x_0'b$

Prediction error: $f_0 = \hat{y}_0 - y_0 = x_0'(b - \beta) + \varepsilon_0$

Some properties of $\hat{y}_0$

- Under assumptions (A1) and (A2), E$\{b\} = \beta$ and $\hat{y}_0$ is an unbiased predictor
- Variance of $\hat{y}_0$ (due to variation of b)

$$V\{\hat{y}_0\} = V\{x_0'b\} = x_0' \, V\{b\} \, x_0 = \sigma^2 x_0'(X'X)^{-1}x_0 = s_0^2$$

- Variance of the prediction error f_0

$$V\{f_0\} = V\{x_0'(b - \beta) + \varepsilon_0\} = \sigma^2(1 + x_0'(X'X)^{-1}x_0) = s_{f0}^2$$

given that ε_0 and b are uncorrelated

Prediction Intervals

$100\gamma\%$ prediction interval

- for the expected value of Y at x_0, i.e., $y_0 = x_0'\beta + \varepsilon_0$: $\hat{y}_0 = x_0'b$

  $$\hat{y}_0 - z_{(1+\gamma)/2}\, s_0 \le y_0 \le \hat{y}_0 + z_{(1+\gamma)/2}\, s_0$$

  with the standard error s_0 of $\hat{y}_0$ from $s_0^2 = \sigma^2\, x_0'(X'X)^{-1}x_0$
- for the prediction Y at x_0

  $$\hat{y}_0 - z_{(1+\gamma)/2}\, s_{f0} \le y_0 \le \hat{y}_0 + z_{(1+\gamma)/2}\, s_{f0}$$

  with s_{f0} from $s_{f0}^2 = \sigma^2\,(1 + x_0'(X'X)^{-1}x_0)$; takes the error term ε_0 into account

Calculation of s_{f0}

- OLS estimate s^2 of σ^2 from regression output (GRETL: “S.E. of regression”)
- Substitution of s^2 for σ^2: $s_0 = s[x_0'(X'X)^{-1}x_0]^{0.5}$, $s_{f0} = [s^2 + s_0^2]^{0.5}$

Example: Simple Regression

Given the relation $y_i = \beta_1 + x_i\beta_2 + \varepsilon_i$

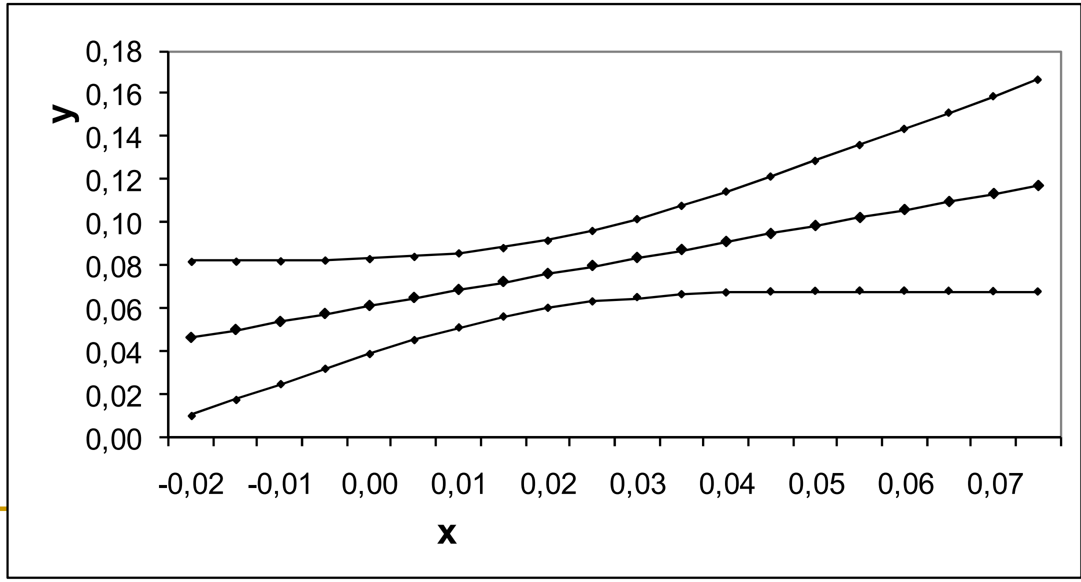
Predictor for Y at x_0, i.e., $y_0 = \beta_1 + x_0\beta_2 + \varepsilon_0$:

$$\hat{y}_0 = b_1 + x_0'b_2$$

Variance of the prediction error

$$V\{\hat{y}_0 - y_0\} = \sigma^2\left(1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{(N-1)s_x^2}\right)$$

Figure: Prediction intervals for various x_0's (indicated as "x") for $\gamma = 0.95$



Individual Wages: Prediction

The fitted model is

$$wage_i = -3.3800 + 1.3444\ male_i + 0.6388\ school_i + 0.1248\ exper_i$$

For a male with *school* = 12 and *exper* = 5, the predicted wage is

$$wage_0 = 6.25405 \approx 6.25$$

Calculation of variance s_0^2:

- Based on variance $s_0^2 = x_0'\, V\{b\}\, x_0 = \sigma^2\, x_0'(X'X)^{-1}x_0$ is laborious
- Re-estimating the model for regressors $m1 = male-1$, $s1 = school-12$, $e1 = exper-5$ gives

  $$wage = 6.25405 + 1.3444\ m1 + 0.6388\ s1 + 0.1248\ e1$$

  with a std.err. of the intercept of 0.10695.
- The std.err. of the intercept, i.e., of the expected wage $wage_0$, is s_0

Individual Wages: Prediction, cont'd

The 95% confidence interval for $wage_0$ is

$$6.25405 - 1.96 * 0.10695 \leq wage_0 \leq 6.25405 + 1.96 * 0.10695$$

or $6.04 \leq wage_0 \leq 6.47$

The 95% prediction interval for $wage_0$:

- From model fit: s = 3.046143
- $s_{f0} = [s^2 + s_0^2]^{0.5} = [3.046143^2 + 0.10695^2]^{0.5} = 3.048$
- 95% prediction interval

$$6.254 - 1.96 * 3.048 \leq wage_0 \leq 6.254 + 1.96 * 3.048$$

or $0.16 \leq wage_0 \leq 12.35$

Your Homework

1. For Verbeek's data set "bwages" use GRETL (a) for estimating a linear regression model with intercept for *wage* p.h. with explanatory variables *male* and *educ*; (b) interpret the coefficients of the model; (c) test the hypothesis that men and women, on average, have the same wage p.h., against the alternative that women's wage p.h. are different from men's wage p.h.; (d) repeat this test against the alternative that women earn less; (e) calculate a 95% confidence interval for the wage difference of males and females.
2. Generate a variable *exper_b* by adding the Binomial random variable $BE \sim B(2,0.5)$ to *exper*; (a) estimate two linear regression models with intercept for *wage* p.h. with explanatory variables (i) *male* and *exper*, and (ii) *male*, *exper_b*, and *exper*; compare the standard errors of the estimated coefficients;

Your Homework

(b) compare the VIFs for the variables of the two models; (c) check the correlations of the involved regressors.

3. The goodness-of-fit statistic R^2 is the portion of the variance in Y that can be explained by the linear regression; $100R^2$ is measured in percent; show that

   $0 \leq R^2 \leq 1$, if the model contains an intercept.

4. Show for a linear regression with intercept that $R^2 >$ adj R^2.