Econometrics - Lecture 3

Regression Models: Interpretation and Comparison

Contents

- The Linear Model: Interpretation
- Selection of Regressors
- Selection Criteria
- Comparison of Competing Models
- Specification of the Functional Form
- Structural Break

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 $Y, X_2, \ldots, X_K$: observable
- Observations: $y_i, x_{i2}, \ldots, x_{iK}$, $i = 1, \ldots, N$
- Error term ε_i (disturbance term) contains all influences that are not included explicitly in the model; unobservable
- Assumption (A1), i.e., $E\{\varepsilon_i \mid X\} = 0$ or $E\{\varepsilon_i \mid x_i\} = 0$, gives

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

  the model describes the expected value of y_i given x_i (conditional expectation)

Example: Wage Equation

Wage equation (Verbeek's dataset "wages1")

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

Answers questions like:

- Expected wage p.h. of a female with 12 years of education and 10 years of experience

Wage equation fitted to all 3294 observations

$$wage_i = -3.38 + 1.34{*}male_i + 0.64{*}school_i + 0.12{*}exper_i$$

- Expected wage p.h. of a female with 12 years of education and 10 years of experience: 5.50 USD

$$wage_i = -3.38 + 1.34{*}0 + 0.64{*}12 + 0.12{*}10 = 5.50$$

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

$$\frac{\Delta E\{y_i \mid x_i\}}{\Delta x_k} = \beta_k \quad \text{for } \Delta x_k = 1$$

- For continuous regressors

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

  Marginal effect of changing X_k on Y

- Ceteris paribus condition: measuring the effect of a change of Y due to a change $\Delta x_k = 1$ by β_k implies
  - knowledge which other X_i, $i \neq k$, are in the model
  - that all other X_i, $i \neq k$, remain unchanged

Example: Coefficients of Wage Equation

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

$$\frac{\partial E\{wage_i \mid male_i, school_i, exper_i\}}{\partial\, school_i} = \beta_3$$

Wage equation fitted to all 3294 observations

$$wage_i = -3.38 + 1.34*male_i + 0.64*school_i + 0.12*exper_i$$

- One extra year at school, e.g., at the university, results in an increase of 64 cents; a 4-year study results in an increase of 2.56 USD of the wage p.h.
- This is true for otherwise (gender, experience) identical people

Regression Coefficients, cont'd

- The marginal effect of a changing regressor may depend on other variables

Examples

- Wage equation: $wage_i = \beta_1 + \beta_2\, male_i + \beta_3\, age_i + \beta_4\, age_i^2 + \varepsilon_i$

  the impact of changing age depends on age:

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

- Wage equation may contain $\beta_3\, age_i + \beta_4\, age_i\, male_i$: marginal effect of age depends upon gender

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

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

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

- For a log-linear model with (log x_i)' = (1, log x_{i2},…, log x_{ik})

  log y_i = (log x_i)' β + ε_i

  elasticities are the coefficients β (see slide 10)

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

Example: Wage Elasticity

Wage equation, fitted to all 3294 observations:

$$\log(wage_i) = 1.09 + 0.20\, male_i + 0.19 \log(exper_i)$$

The coefficient of $\log(exper_i)$ measures the elasticity of wages with respect to experience:

- 100% more years of experience result in an increase of wage by 0.19 or a 19% higher wage
- 10% more years of experience result in a 1.9% higher wage

Elasticities, continues slide 8

This follows – for log y_i = (log x_i)' β + ε_i – from

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

$$\approx \frac{\partial \log E\{y_i | x_i\}}{\partial E\{y_i | x_i\}} \frac{\partial E\{y_i | x_i\}}{\partial x_{ik}} = \frac{1}{E\{y_i | x_i\}} \frac{\partial E\{y_i | x_i\}}{\partial x_{ik}}$$

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

and

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

$$= \frac{\beta_k}{x_{ik}} x_{ik} = \beta_k$$

Semi-Elasticities

Semi-elasticity: measures the *relative* change in the dependent variable *Y* due to an (absolute) one-unit-change in X_k

- Linear regression for

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

  the elasticity of *Y* with respect to X_k is

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

  β_k measures the relative change in *Y* due to a change in X_k by one unit

- β_k is called semi-elasticity of *Y* with respect to X_k

Example: Wage Differential

Wage equation, fitted to all 3294 observations:

$$log(wage_i) = 1.09 + 0.20\ male_i + 0.19 \log(exper_i)$$

- The semi-elasticity of the wages with respect to gender, i.e., the relative wage differential between males and females, is the coefficient of $male_i$: 0.20 or 20%
- The wage differential between males ($male_i$ =1) and females is obtained from $wage_f = \exp\{1.09 + 0.19 \log(exper_i)\}$ and $wage_m = wage_f \exp\{0.20\} = 1.22\ wage_f$; the wage differential is 0.22 or 22%; the coefficient 0.20[1] is a good approximation.

---

[1] For small x, $\exp\{x\} = \Sigma_k x^k/k! \approx 1+x$

Contents

- The Linear Model: Interpretation
- Selection of Regressors
- Selection Criteria
- Comparison of Competing Models
- Specification of the Functional Form
- Structural Break

Selection of Regressors

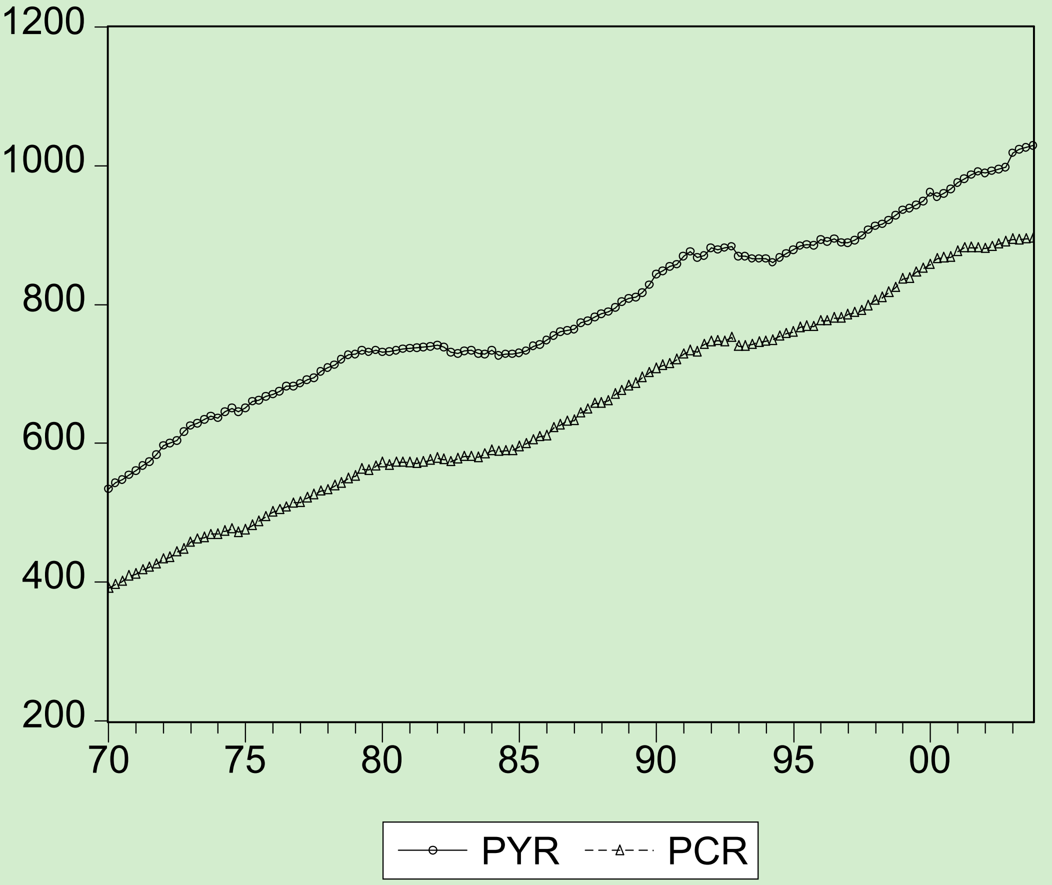
Specification errors:

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

Questions:

- What are the consequences of a specification error?
- How to avoid specification errors?
- How to detect an erroneous specification?

Example: Income and Consumption



PCR: Private Consumption, real, in bn. EUROs

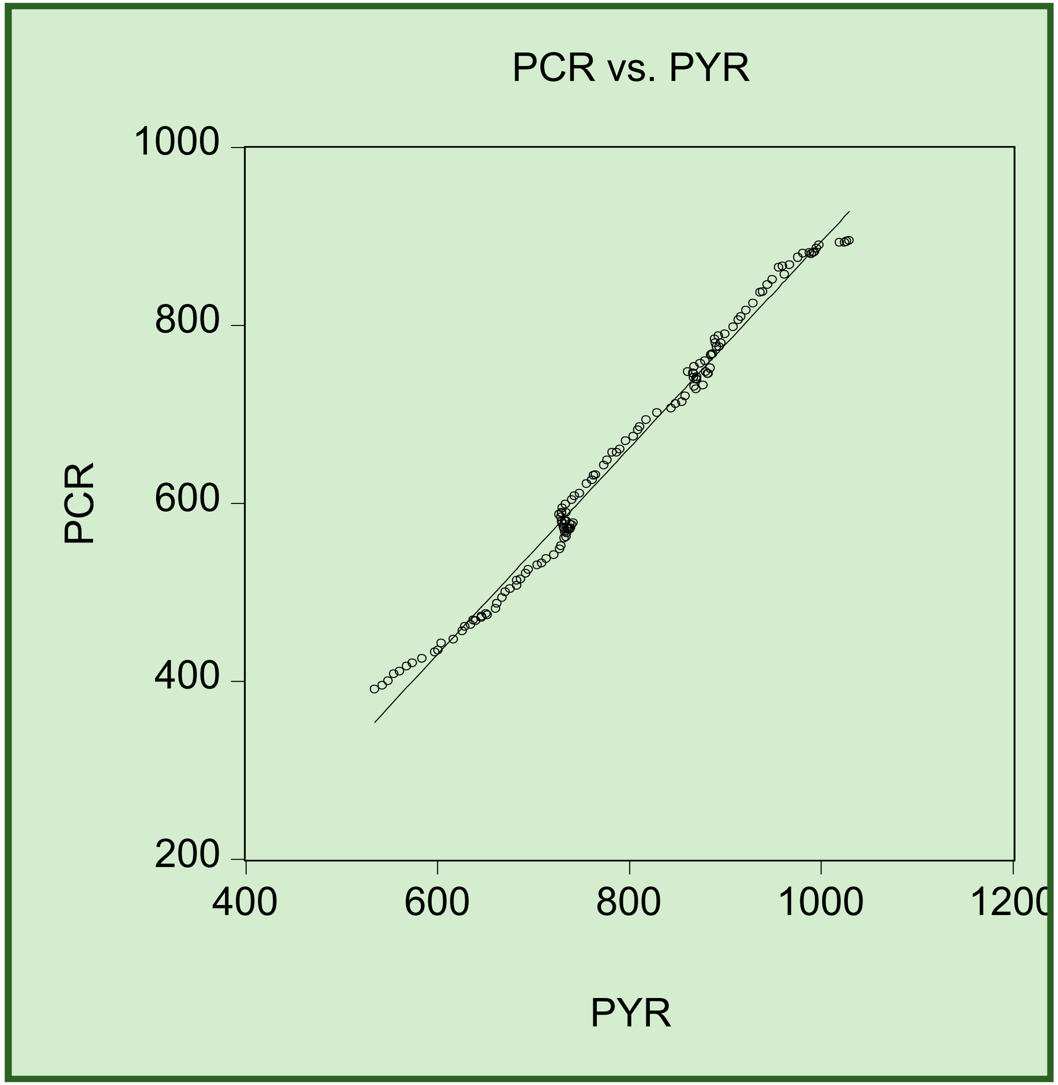
PYR: Household's Disposable Income, real, in bn. EUROs

1970:1-2003:4

Basis: 1995

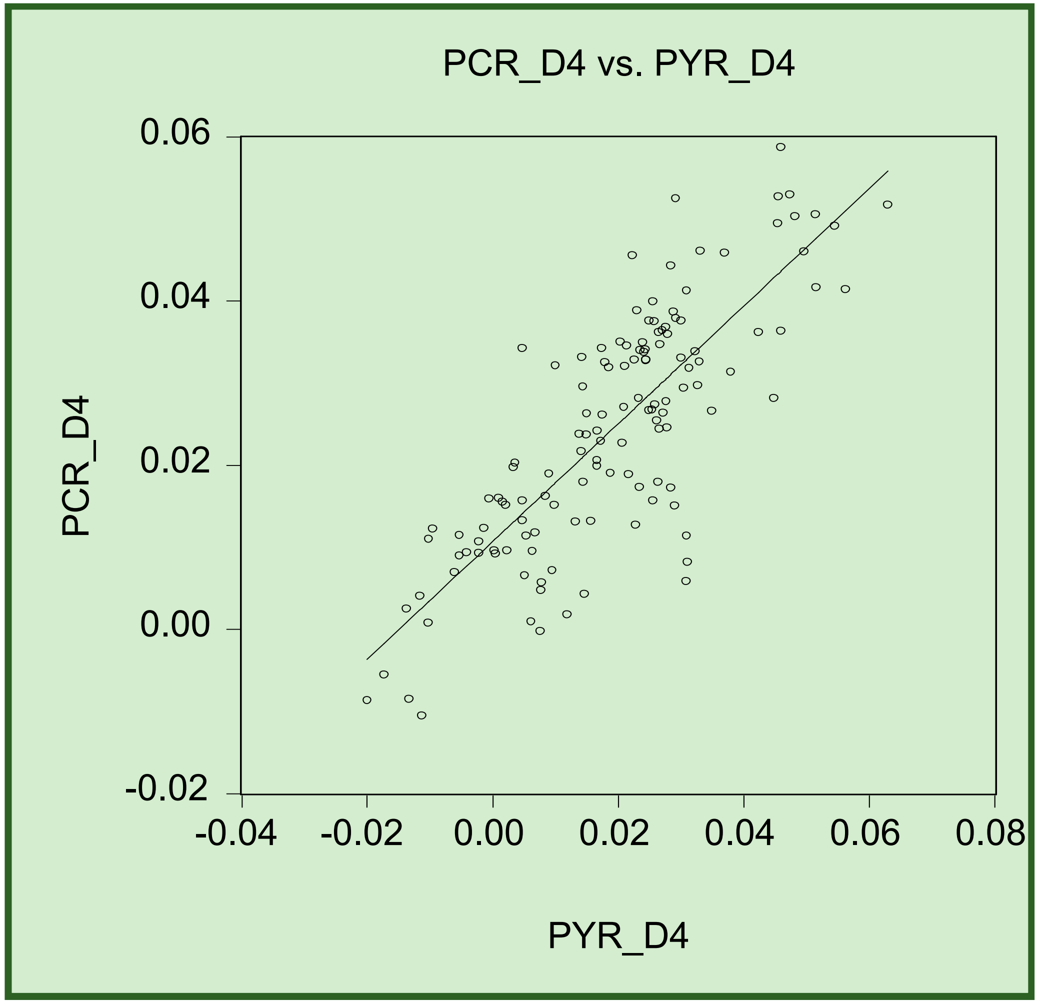
Source: AWM-Database

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, yearly growth rate

PYR_D4: Household's Disposable Income, real, yearly growth rate

1970:1-2003:4

Basis: 1995

Source: AWM-Database

Consumption Function

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

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

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

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

Consumption function with trend T_i = i/1000:

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

Consumption Function, cont’d

OLS estimated consumption function: Output from GRETL

```
Dependent variable : PCR_D4

              coefficient          std. error           t-ratio          p-value
  ------------------------------------------------------------
  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    ***

Mean dependent var                 0,024911             S.D. dependent var           0,015222
Sum squared resid                  0,007726              S.E. of regression          0,007739
R- squared                         0,745445              Adjusted R-squared          0,741498
F(2, 129)                          188,8830              P-value (F)                  4,71e-39
Log-likelihood                     455,9302             Akaike criterion            -905,8603
Schwarz criterion                 -897,2119             Hannan-Quinn                -902,3460
rho                                0,701126             Durbin-Watson                0,601668
```

Misspecification: Two Models

Two models:

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

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

with J-vector z_i

Misspecification: Omitted Regressor

Specified model is (B), but true model is (A)

$y_i = x_i'\beta + z_i'\gamma + \varepsilon_i$ (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., J 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

Specified model is (A), but true model is (B):

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

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

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 higher variances and standard errors than the OLS estimate b_B obtained from fitting model (B)

Consequences

Consequences of specification errors:

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

Specification Search

General-to-specific modeling:

1. List all potential regressors, based on, e.g.,
    - economic theory
    - empirical research
    - availability of data
2. Specify the most general model: include all potential regressors
3. Iteratively, test which variables have to be dropped, re-estimate
4. Stop if no more variable has to be dropped

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

Specification Search, cont'd

Alternative procedures

- Specific-to-general modeling: start with a small model and add variables as long as they contribute to explaining Y
- Stepwise regression

Specification search can be subsumed under *data mining*

Practice of Specification Search

Applied research

- Starts with a – in terms of economic theory – plausible specification
- Tests whether imposed restrictions are correct, such as
  - Test for omitted regressors
  - Test for autocorrelation of residuals
  - Test for heteroskedasticity
- Tests whether further restrictions need to be imposed
  - Test for irrelevant regressors

Obstacles for good specification

- Complexity of economic theory
- Limited availability of data

Contents

- The Linear Model: Interpretation
- Selection of Regressors
- Selection Criteria
- Comparison of Competing Models
- Specification of the Functional Form
- Structural Break

Regressor Selection Criteria

Criteria for adding and deleting regressors

- *t*-statistic, *F*-statistic
- Adjusted R^2
- Information Criteria: penalty for increasing number of regressors
  - Akaike's Information Criterion

$$AIC = \log \tfrac{1}{N} \sum_i e_i^2 + \tfrac{2K}{N}$$

  - Alternative criteria are
    - Schwarz's Bayesian Information Criterion (BIC)
    - Hannan-Quinn Information Criterion

The model with relevant regressors, with higher adj R^2, the smaller AIC is preferred

Information Criteria

The most popular information criteria are

- Akaike's Information Criterion

$$AIC = \log \tfrac{1}{N} \sum_i e_i^2 + \tfrac{2K}{N}$$

- Schwarz's Bayesian Information Criterion

$$BIC = \log \tfrac{1}{N} \sum_i e_i^2 + \tfrac{K}{N} \log N$$

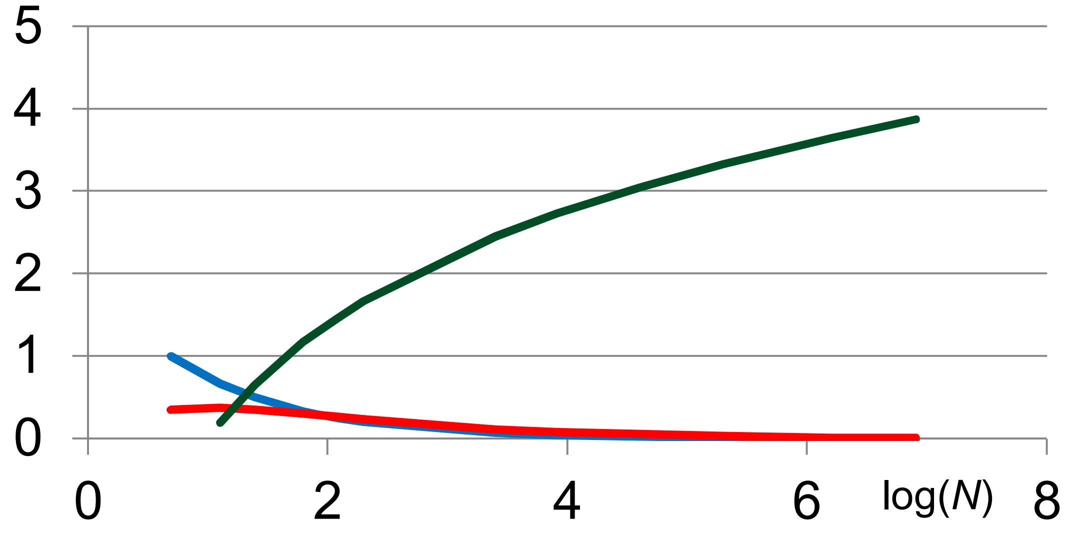
- Hannan-Quinn Information Criterion

$$HQIC = \log \tfrac{1}{N} \sum_i e_i^2 + 2K \log \log N$$

Decide in favour of the model with the *lowest* value of the information criterion

Information Criteria: Penalties

- Akaike
  $2/N$
- Schwarz
  $\log(N)/N$
- Hannan-Quinn
  $2\log(\log(N))$



N	$\log(N)$	AIC	BIC	HQC
2	0,69	1,00	0,35	-0,73
3	1,10	0,67	0,37	0,19
4	1,39	0,50	0,35	0,65
6	1,79	0,33	0,30	1,17
8	2,08	0,25	0,26	1,46
10	2,30	0,20	0,23	1,67
30	3,40	0,07	0,11	2,45
50	3,91	0,04	0,08	2,73
100	4,61	0,02	0,05	3,05
200	5,30	0,01	0,03	3,33
500	6,21	0,00	0,01	3,65
1000	6,91	0,00	0,01	3,87

Wages: Which Regressors?

Are *school* and *exper* relevant regressors in

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

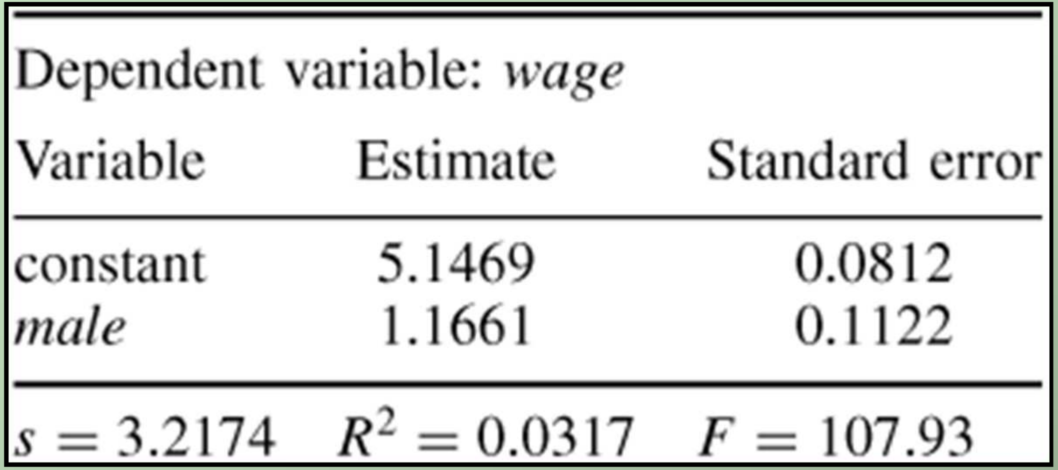
or shall they be omitted?

- t-test: p-values are 4.62E-80 (*school*) and 1.59E-7 (*exper*)
- F-test: $F = [(0.1326-0.0317)/2]/[(1-0.1326)/(3294-4)] = 191.24$, with p-value 2.68E-79
- adj R^2: 0.1318 for the wider model, much higher than 0.0315
- AIC: the wider model (AIC = 16690.2) is preferable; for the smaller model: AIC = 17048.5
- BIC: the wider model (BIC = 16714.6) is preferable; for the smaller model: BIC = 17060.7

All criteria suggest the wider model

Wages, cont’d

OLS estimated smaller wage equation (Table 2.1, Verbeek)



Dependent variable: *wage*

Variable	Estimate	Standard error
constant	5.1469	0.0812
male	1.1661	0.1122

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

with AIC = 17048.46, BIC = 17060.66

Wages, cont’d

OLS estimated wider wage equation (Table 2.2, Verbeek)

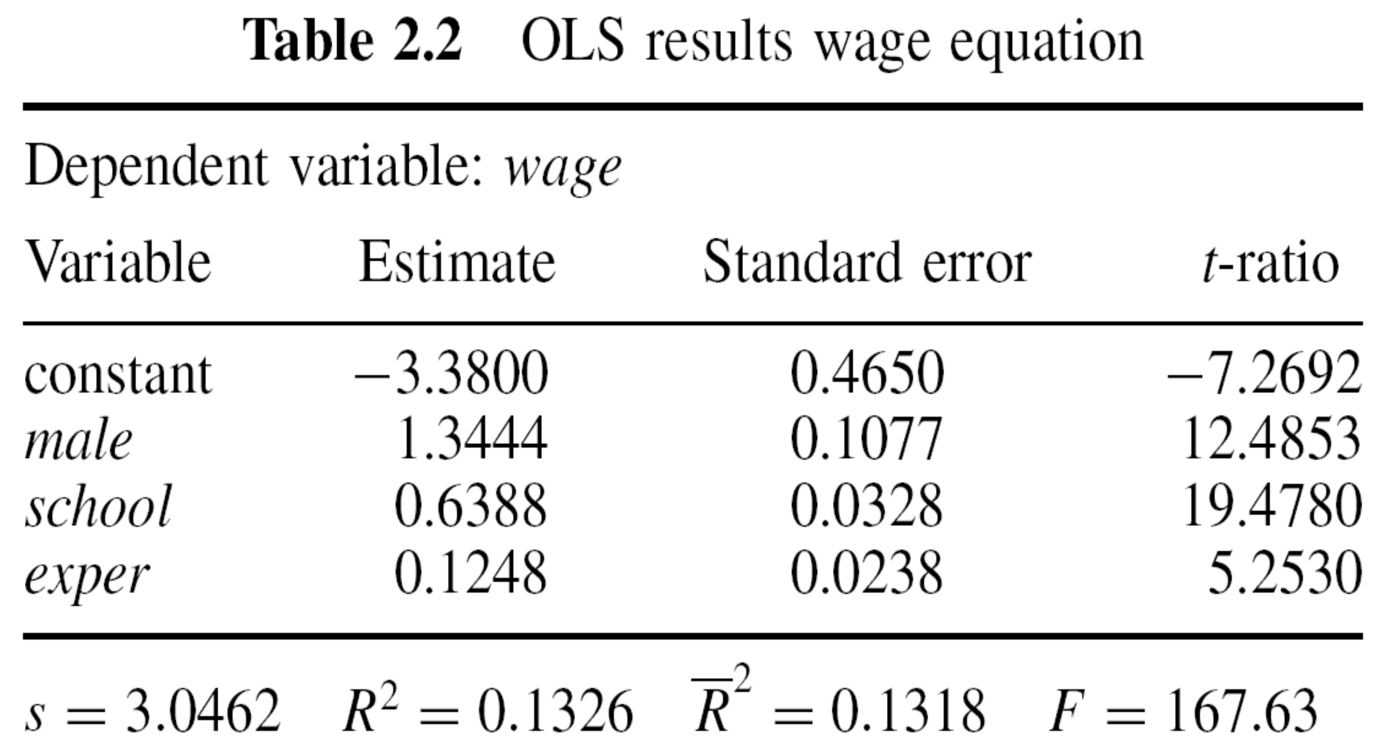


Table 2.2 OLS results wage equation

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

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

with AIC = 16690.18, BIC = 16714.58

The AIC Criterion

Various versions in literature

- Verbeek, also Greene:

$$AIC_V = \log \tfrac{1}{N} \sum\nolimits_i e_i^2 + \tfrac{2K}{N} = \log(s^2) + 2K / N$$

- Akaike's original formula is

$$AIC_A = -2\, l(b)/N + 2K/N = AIC_V + 1 + \log(2\pi)$$

with the log-likelihood function

$$\ell(b) = -\frac{N}{2}\left(1 + \log(2\pi) + \log(s^2)\right)$$

- GRETL:

$$AIC_G = N \log(s^2) + 2K + N\left(1 + \log(2\pi)\right) = N\, AIC_A$$

Contents

- The Linear Model: Interpretation
- Selection of Regressors
- Selection Criteria
- Comparison of Competing Models
- Specification of the Functional Form
- Structural Break

Nested Models: Comparison

Model (B), $y_i = x_i'\beta + v_i$, see slide 21, is nested in model

$$y_i = x_i'\beta + z_i'\gamma + \varepsilon_i \qquad \text{(A)}$$

i.e., (A) is extended by J additional regressors z_i

Do the J added regressors contribute to explaining Y?

- F-test (t-test when $J = 1$) for testing H_0: all coefficients of added regressors are zero

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

  R_B^2 and R_A^2 are the R^2 of the models without (B) and with (A) the J additional regressors, respectively
- Adjusted R^2: adj R_A^2 > adj R_B^2 equivalent to $F > 1$
- Information Criteria: choose the model with the smaller value of the information criterion

Comparison of Non-nested Models

Non-nested models:

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

$y_i = z_i'\gamma + v_i$ (B)

at least one component in z_i that is not in x_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}: regressors from z_i not in x_i

  $y_i = z_i'\gamma + x_{2i}'\delta_A + v^*_i$ with x_{2i}: regressors from x_i not in z_i

  - Test validity of model A by testing H_0: $\delta_B = 0$
  - Analogously, test validity of model B by testing H_0: $\delta_A = 0$
  - Possible results: A or B is valid, both models are valid, none is valid
- Other procedures: *J*-test, PE-test (see below)

Wages: Which Model?

Which of the models is adequate?

$$\log(wage_i) = 0.119 + 0.260\, male_i + 0.115\, school_i \qquad \text{(A)}$$

adj $R^2 = 0.121$, BIC = 5824.90,

$$\log(wage_i) = 0.119 + 0.064\, age_i \qquad \text{(B)}$$

adj $R^2 = 0.069$, BIC = 6004.60

- Artificially nested model

$$\log(wage_i) =$$

$$= -0.472 + 0.243\, male_i + 0.088\, school_i + 0.035\, age_i$$

- Test of model validity
  - model A: *t*-test for *age*, *p*-value 5.79E-15; model A is not adequate
  - model B: *F*-test for *male* and *school*: model B is not adequate

J-Test: Comparison of Non-nested Models

Non-nested models: (A) $y_i = x_i'\beta + \varepsilon_i$, (B) $y_i = z_i'\gamma + v_i$ with components of z_i that are not in x_i

- Combined model

$$y_i = (1 - \delta)\, x_i'\beta + \delta\, z_i'\gamma + u_i$$

with $0 < \delta < 1$; δ indicates model adequacy

- Transformed model

$$y_i = x_i'\beta^* + \delta z_i'c + u_i = x_i'\beta^* + \delta \hat{y}_{iB} + u^*_i$$

with OLS estimate c for γ and predicted values $\hat{y}_{iB} = z_i'c$ obtained from fitting model B; $\beta^* = (1-\delta)\beta$

- *J*-test for validity of model A by testing H_0: $\delta = 0$
- Less computational effort than the encompassing *F*-test

Wages: Which Model?

Which of the models is adequate?

$$\log(wage_i) = 0.119 + 0.260\, male_i + 0.115\, school_i \qquad \text{(A)}$$

adj $R^2 = 0.121$, BIC = 5824.90,

$$\log(wage_i) = 0.119 + 0.064\, age_i \qquad \text{(B)}$$

adj $R^2 = 0.069$, BIC = 6004.60

Test the validity of model B by means of the *J*-test

- Extend the model B to

  $$\log(wage_i) = -0.587 + 0.034\, age_i + 0.826\, \hat{y}_{iA}$$

  with values $\hat{y}_{iA}$ predicted for log(*wage*$_i$) from model A
- Test of model validity: *t*-test for coefficient of $\hat{y}_{iA}$, $t = 15.96$, *p*-value 2.65E-55
- Model B is not a valid model

Linear vs. Log-linear Model

Choice between linear and log-linear functional form

$$y_i = x_i'\beta + \varepsilon_i \qquad \text{(A)}$$

$$\log y_i = (\log x_i)'\beta + v_i \qquad \text{(B)}$$

- In terms of economic interpretation: Are effects additive or multiplicative?
- Log-transformation stabilizes variance, particularly if the dependent variable has a skewed distribution (wages, income, production, firm size, sales,…)
- Log-linear models are easily interpretable in terms of elasticities

PE-Test: Linear vs. Log-linear Model

Choice between linear and log-linear functional form

- Estimate both models

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

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

  calculate the fitted values y_f (from model A) and log y_f (from B)
- Test $H_0: \delta_{LIN} = 0$ in

  $y_i = x_i'\beta + \delta_{LIN}\,(\log (y\_f_i) - \log y\_f_i) + u_i$

  not rejecting $H_0: \delta_{LIN} = 0$ favors the model A
- Test $H_0: \delta_{LOG} = 0$ in

  $\log y_i = (\log x_i)'\beta + \delta_{LOG}\,(y\_f_i - \exp\{\log y\_f_i\}) + u_i$

  not rejecting $H_0: \delta_{LOG} = 0$ favors the model B
- Both null hypotheses are rejected: find a more adequate model

Wages: Which Model?

Test of validity of models by means of the PE-test

The fitted models are (with l_x for log(x))

$$wage_i = -2.046 + 1.406\ male_i + 0.608\ school_i \qquad (A)$$

$$l\_wage_i = 0.119 + 0.260\ male_i + 0.115\ l\_school_i \qquad (B)$$

- x_f: predicted value of x: d_log = log($wage\_f$) – l_wage_f, d_lin = $wage\_f$ – exp(l_wage_f)
- Test of validity of model A:

  $$wage_i = -1.708 + 1.379\ male_i + 0.637\ school_i - 4.731\ d\_log_i$$

  with p-value 0.013 for d_log; validity of model A in doubt
- Test of model validity, model B:

  $$l\_wage_i = -1.132 + 0.240\ male_i + 1.008\ l\_school_i + 0.171\ d\_lin_i$$

  with p-value 0.076 for d_lin; model B to be preferred

The PE-Test

Choice between linear and log-linear functional form

- The auxiliary regressions are estimated for testing purposes
- If the linear model is not rejected: accept the linear model
- If the log-linear model is not rejected: accept the log-linear model
- If both are rejected, neither model is appropriate, a more adequate model should be considered
- In case of the Individual Wages example:
  - Linear model (A): *t*-statistic is – 4.731, *p*-value 0.013: the model is rejected
  - Log-linear model (B): *t*-statistic is 0.171, *p*-value 0.076 : the model is not rejected

Contents

- The Linear Model: Interpretation
- Selection of Regressors
- Selection Criteria
- Comparison of Competing Models
- Specification of the Functional Form
- Structural Break

Non-linear Functional Forms

Model specification

$$y_i = g(x_i, \beta) + \varepsilon_i$$

substitution of $g(x_i, \beta)$ for $x_i'\beta$: allows for two types on non-linearity

- $g(x_i, \beta)$ non-linear in regressors (but linear in coefficients)
  - Powers of regressors, e.g., $g(x_i, \beta) = \beta_1 + \beta_2\, age_i + \beta_3\, age_i^2$
  - Interactions of regressors, e.g., $g(x_i, \beta) = \beta_1 + \beta_2\, age_i + \beta_3\, age_i{*}male_i$

  OLS technique still works; *t*-test, *F*-test for specification check
- $g(x_i, \beta)$ non-linear in regression coefficients, e.g.,
  - $g(x_i, \beta) = \beta_1\, x_{i1}^{\beta 2}\, x_{i2}^{\beta 3}$

    logarithmic transformation: $\log g(x_i, \beta) = \log \beta_1 + \beta_2 \log x_{i1} + \beta_3 \log x_{i2}$
  - $g(x_i, \beta) = \beta_1 + \beta_2\, x_i^{\beta 3}$

    non-linear least squares estimation, numerical procedures

  Various specification test procedures, e.g., RESET test, Chow test

Individual Wages: Effect of Gender and Education

Effect of gender may be depending of education level

- Separate models for males and females
- Interaction terms between dummies for education level and male

Example: Belgian Household Panel, 1994 (“bwages”, N=1472)

- Five education levels
- Model for log(*wage*) with education dummies; see next slide
- Model with interaction terms between education dummies and gender dummy; see slide 49
- F-statistic for interaction terms:

  $$F(5, 1460) = \{(0.4032-0.3976)/5\}/\{(1-0.4032)/(1472-12)\}$$
  $$= 2.74$$

  with a p-value of 0.018

Wages: Model with Education Dummies

Model with education dummies: Verbeek, Table 3.11

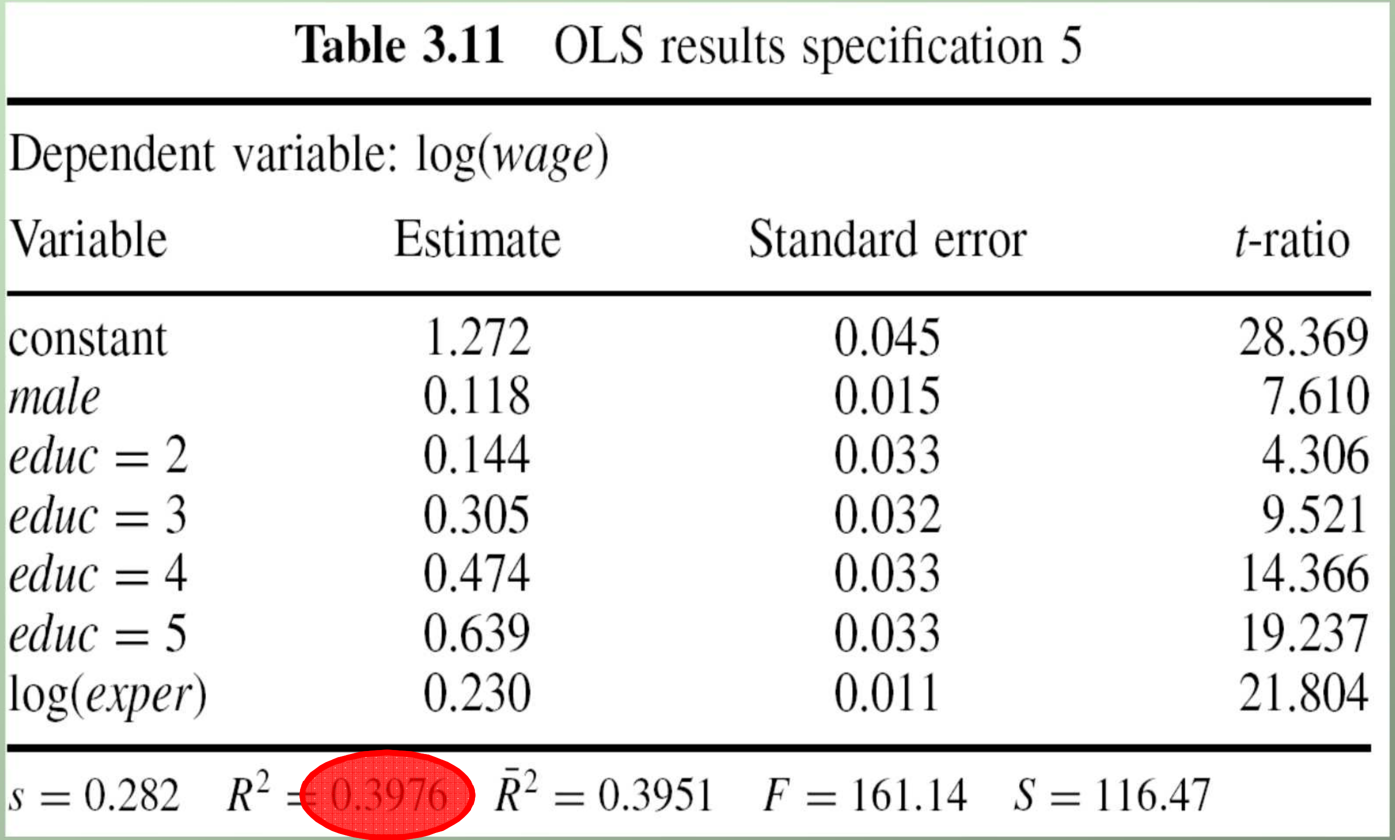


Table 3.11 OLS results specification 5

Dependent variable: log(*wage*)

Variable	Estimate	Standard error	t-ratio
constant	1.272	0.045	28.369
male	0.118	0.015	7.610
$educ = 2$	0.144	0.033	4.306
$educ = 3$	0.305	0.032	9.521
$educ = 4$	0.474	0.033	14.366
$educ = 5$	0.639	0.033	19.237
log(*exper*)	0.230	0.011	21.804

$s = 0.282 \quad R^2 = 0.3976 \quad \bar{R}^2 = 0.3951 \quad F = 161.14 \quad S = 116.47$

Wages: Model with Gender Interactions

Wage equation with interactions *educ*male*

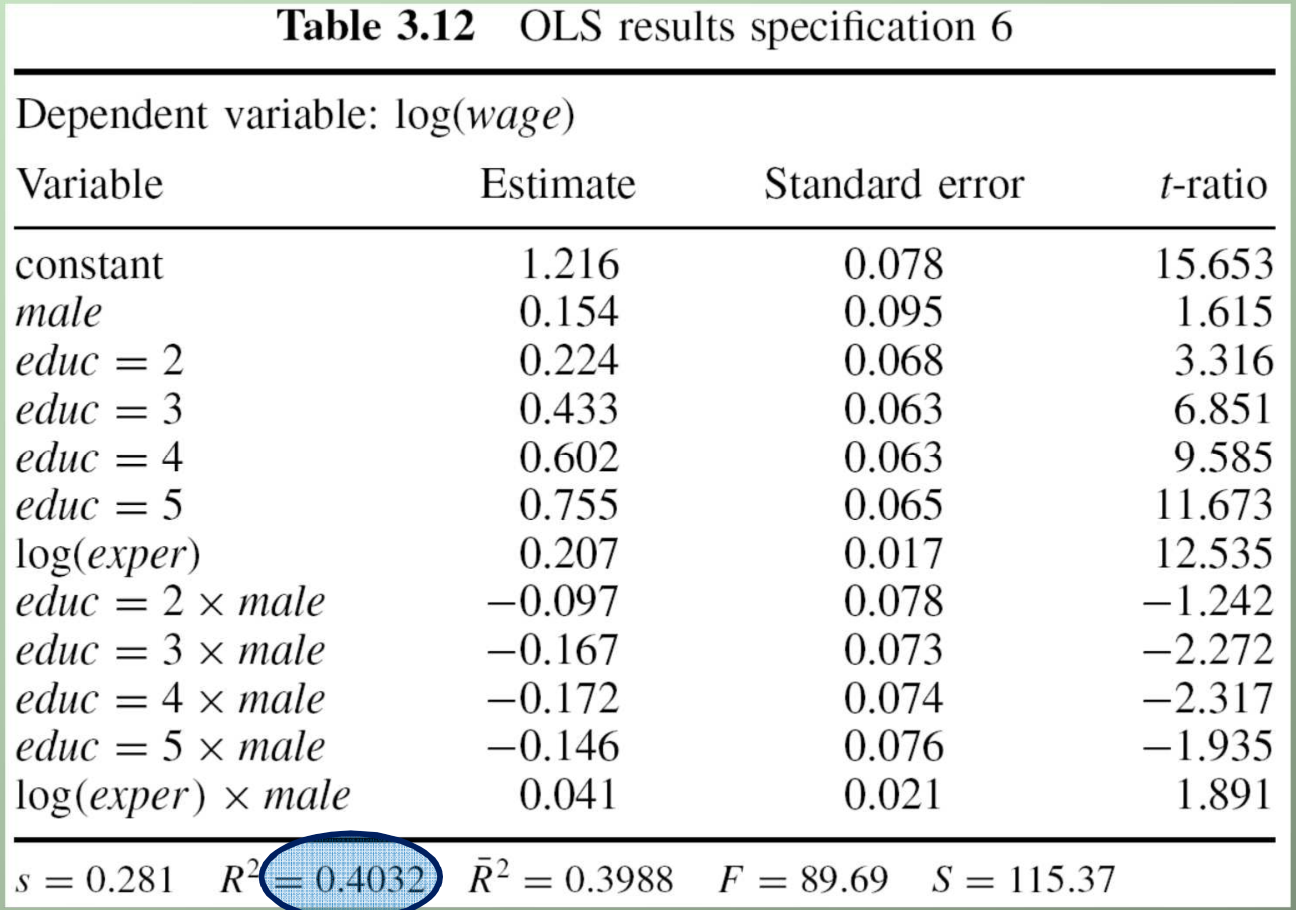


Table 3.12 OLS results specification 6

Dependent variable: log(*wage*)

Variable	Estimate	Standard error	t-ratio
constant	1.216	0.078	15.653
male	0.154	0.095	1.615
$educ = 2$	0.224	0.068	3.316
$educ = 3$	0.433	0.063	6.851
$educ = 4$	0.602	0.063	9.585
$educ = 5$	0.755	0.065	11.673
$\log(exper)$	0.207	0.017	12.535
$educ = 2 \times male$	−0.097	0.078	−1.242
$educ = 3 \times male$	−0.167	0.073	−2.272
$educ = 4 \times male$	−0.172	0.074	−2.317
$educ = 5 \times male$	−0.146	0.076	−1.935
$\log(exper) \times male$	0.041	0.021	1.891

$s = 0.281$ $R^2 = 0.4032$ $\bar{R}^2 = 0.3988$ $F = 89.69$ $S = 115.37$

RESET Test

Test of the linear model E{y_i |x_i} = x_i'β against misspecification of the functional form:

- Null hypothesis: linear model is correct functional form
- Test of H_0: RESET test (Regression Specification Error Test), Ramsey (1969)
- Test idea: linear model is extended by adding $\hat{y}_i^2$, $\hat{y}_i^3$, ..., where $\hat{y}_i$ is the fitted values from the linear model; extension does not improve model fit under H_0
  - $\hat{y}_i^2$ is a function of squares (and interactions) of the regressor variables; analogously for $\hat{y}_i^3$, ...
  - If the *F*-test indicates that the additional regressor $\hat{y}_i^2$ contributes to explaining *Y*: the linear relation is not adequate, another functional form is more appropriate

The RESET Test Procedure

Test of the linear model $E\{y_i \mid x_i\} = x_i'\beta$ against misspecification of the functional form:

- Linear model extended by adding $\hat{y}_i^2, \ldots, \hat{y}_i^Q$
- F- (or t-) test to decide whether $\hat{y}_i^2, \ldots, \hat{y}_i^Q$ contribute as additional regressors to explaining Y
- Maximal power Q of fitted values: typical choice is $Q = 2$ or $Q = 3$

In **GRETL**: Ordinary Least Squares… => Tests => Ramsey's RESET, input of Q

Wages: RESET Test

The fitted models are (with *l_x* for log(*x*))

$$wage_{\mathrm{i}} = -2.046 + 1.406\ male_{\mathrm{i}} + 0.608\ school_{\mathrm{i}} \qquad \text{(A)}$$

$$l\_wage_{\mathrm{i}} = 0.119 + 0.260\ male_{\mathrm{i}} + 0.115\ l\_school_{i} \qquad \text{(B)}$$

Test of specification of the functional form with $Q = 3$

- Model A: Test statistic: $F(2, 3288) = 10.23$, *p*-value = 3.723e-005
- Model B: Test statistic: $F(2, 3288) = 4.52$, *p*-value = 0.011

For both models the adequacy of the functional form is in doubt

Contents

Structural Break: Chow Test

In time-series context, coefficients of a model may change due to a major policy change, e.g., the oil price shock

- Modeling a process with structural break

  $$E\{y_i \mid x_i\} = x_i'\beta + g_i x_i'\gamma$$

  with dummy variable $g_i=0$ before the break, $g_i=1$ after the break
- Regressors x_i, coefficients β before, β+γ after the break
- Null hypothesis: no structural break, $\gamma=0$
- Test procedure: fitting the extended model, *F*- (or *t*-) test of $\gamma=0$

  $$F = \frac{S_r - S_u}{S_u} \frac{N-2K}{K}$$

  with S_r (S_u): sum of squared residuals of the (un)restricted model
- Chow test for structural break or structural change, Chow (1960)

Chow Test: The Practice

Test procedure is performed in the following steps

- Fit the restricted model: S_r
- Fit the extended model: S_u
- Calculate F and the p-value from the F-distribution with K and $N-2K$ d.f.

Needs knowledge of break point

In **GRETL**: Ordinary Least Squares… => Tests => Chow test

input of the first observation period after the break point

Your Homework

1. Use the data set “bwages” of Verbeek for the following analyses:
   a) Estimate the model where the hourly wages (*wage*) are explained by *exper*, *male*, and *educ*; interpret the results.
   b) Educ represents the level of education; what does that mean for the estimated coefficient for *educ* in task a).
   c) Repeat task a) using dummy variables for the education levels, *d2* for *educ* = 2, …, *d5* for *educ* = 5 instead of the variable *educ*; compare the models from this and task a) by using (i) the non-nested *F*-test and (ii) the *J*-test; interpret the results.
   d) Use the PE-test to decide whether the model of a) (where hourly wages, *wage,* are explained) or the same model but with *lnwage,* log hourly wages, as explained variable is to be preferred; interpret the result.
   e) Estimate the model for log hourly wages (*lnwage*) with regressors *lnexper*, *male*, *educ,* and the interaction *male*lnexper* as additional regressor; interpret the results.

Your Homework, cont'd

2. OLS is used to estimate β from $y_i = x_i'\beta + \varepsilon_i$, but a relevant regressor z_i is neglected: $y_i = x_i'\beta + z_i'\gamma + \varepsilon_i$. (a) Show that the estimate b is biased, and derive an expression for the bias; (b) what test statistic can be used for testing H_0: $\gamma = 0$?
3. The linear regression is specified as

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

   Show that the elasticity of Y with respect to X_k is $\beta_k x_{ik}$.