ES1004 Econometrics by Example

Lecture 7: Model Specification Errors

Dr. Hany Abdel-Latif

Swansea University, UK

Gujarati textbook, second edition

23rd July 2016

CLRM Assumptions

A$_1$: model is linear in parameters

A$_2$: regressors are fixed non-stochastic

A$_3$: the expected value of the error term is zero $E(u_i|X) = 0$

A$_4$: homoscedastic or constant variance of errors $var(u_i|X) = \sigma^2$

A$_5$: no autocorrelation, $cov(u_i, u_j) = 0, i \neq j$

A$_6$: no multicollinearity; no perfect linear relationships among the Xs

A$_7$: no specification bias

Basic Idea I

- CLRM assumes the model is 'correctly' specified
  - there is no such thing as a perfect model
- an econometric model tries to capture the main features of an economic phenomenon
  - taking into account the underlying economic theory, prior empirical work, intuition, and research skills
- no model can take into account every single factor that affects a particular object of research

Basic Idea II

- a ‘correctly’ specified model . . .
  1. does not exclude any "core" variables
  2. does not include superfluous variables
  3. has the suitable functional form
  4. has no measurement errors

Basic Idea III

- a ‘correctly’ specified model . . .

  5. takes into account outliers in the data
  6. the probability distribution of the error term is well specified
  7. includes non-stochastic regressors
  8. no simultaneity bias

Example: Wage Determination

- a model of hourly wage determination
- data `table1.1`
- CPS: current population survey 1995 data on 1289 workers

Determinants of hourly wage rate

Dependent Variable: WAGE
Method: Least Squares
Date: 07/23/16 Time: 12:18
Sample: 1 1289
Included observations: 1289

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-7.183338	1.015788	-7.071691	0.0000
FEMALE	-3.074875	0.364616	-8.433184	0.0000
NONWHITE	-1.565313	0.509188	-3.074139	0.0022
UNION	1.095976	0.506078	2.165626	0.0305
EDUCATION	1.370301	0.065904	20.79231	0.0000
EXPER	0.166607	0.016048	10.38205	0.0000

R-squared	0.323339	Mean dependent var	12.36585
Adjusted R-squared	0.320702	S.D. dependent var	7.896350
S.E. of regression	6.508137	Akaike info criterion	6.588627
Sum squared resid	54342.54	Schwarz criterion	6.612653
Log likelihood	-4240.370	Hannan-Quinn criter.	6.597646
F-statistic	122.6149	Durbin-Watson stat	1.897513
Prob(F-statistic)	0.000000		

Determinants of hourly wage rate

Dependent Variable: WAGE
Method: Least Squares
Date: 07/23/16 Time: 12:20
Sample: 1 1289
Included observations: 1289

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-8.419035	1.035710	-8.128758	0.0000
FEMALE	-3.009360	0.361432	-8.326210	0.0000
NONWHITE	-1.536077	0.504448	-3.045066	0.0024
UNION	1.026979	0.501521	2.047728	0.0408
EDUCATION	1.323745	0.065937	20.07597	0.0000
EXPER	0.424463	0.053580	7.922075	0.0000
EXPER^2	-0.006183	0.001227	-5.039494	0.0000

R-squared	0.336483	Mean dependent var	12.36585
Adjusted R-squared	0.333378	S.D. dependent var	7.896350
S.E. of regression	6.447128	Akaike info criterion	6.570563
Sum squared resid	53286.93	Schwarz criterion	6.598593
Log likelihood	-4227.728	Hannan-Quinn criter.	6.581084
F-statistic	108.3548	Durbin-Watson stat	1.901169
Prob(F-statistic)	0.000000		

Determinants of hourly wage rate

Dependent Variable: WAGE
Method: Least Squares
Date: 07/23/16 Time: 12:21
Sample: 1 1289
Included observations: 1289

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-9.200668	1.072115	-8.581791	0.0000
FEMALE	-1.433980	0.680797	-2.106326	0.0354
NONWHITE	-1.481891	0.503577	-2.942731	0.0033
UNION	0.949027	0.501081	1.893958	0.0585
EDUCATION	1.318365	0.065801	20.03554	0.0000
EXPER	0.471974	0.056212	8.396344	0.0000
EXPER^2	-0.006274	0.001224	-5.124559	0.0000
EXPER*FEMALE	-0.084151	0.030848	-2.727939	0.0065

R-squared	0.340315	Mean dependent var	12.36585
Adjusted R-squared	0.336711	S.D. dependent var	7.896350
S.E. of regression	6.430992	Akaike info criterion	6.566322
Sum squared resid	52979.16	Schwarz criterion	6.598357
Log likelihood	-4223.994	Hannan-Quinn criter.	6.578346
F-statistic	94.40528	Durbin-Watson stat	1.892702
Prob(F-statistic)	0.000000		

Ramsey's RESET test I

1. from the 'incorrectly' estimated model, obtain the estimated, or fitted, values of the dependent variable
2. reestimate the original model including the fitted values of the dependent variable as additional regressors
3. the initial model is the restricted model and the model is step 2 is the unrestricted model

The Economic Society

Ramsey's RESET test II

4. under the null hypothesis that the restricted (i.e., the original) model is correct, we can use the F test
5. if the F test in step 4 is statistically significant, we can reject the null hypothesis.
    - the restricted model is not appropriate in the present situation

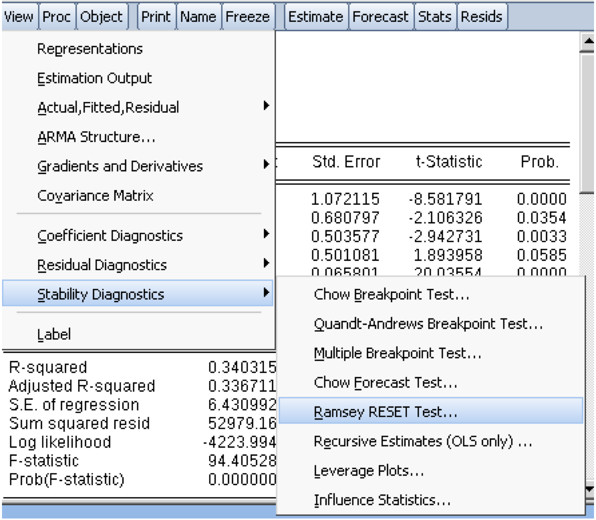
An example: Restricted model

Dependent Variable: WAGE
Method: Least Squares
Date: 07/23/16 Time: 12:18
Sample: 1 1289
Included observations: 1289

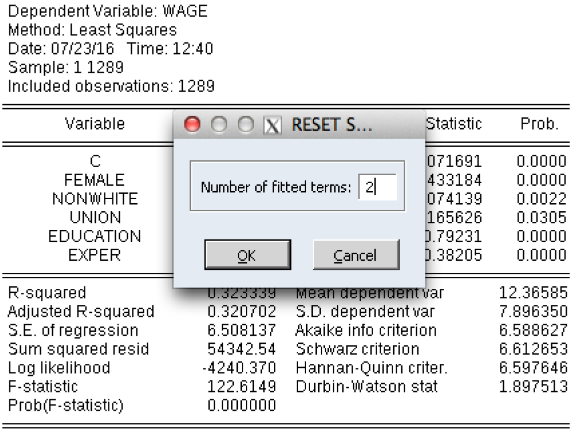
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-7.183338	1.015788	-7.071691	0.0000
FEMALE	-3.074875	0.364616	-8.433184	0.0000
NONWHITE	-1.565313	0.509188	-3.074139	0.0022
UNION	1.095976	0.506078	2.165626	0.0305
EDUCATION	1.370301	0.065904	20.79231	0.0000
EXPER	0.166607	0.016048	10.38205	0.0000

R-squared	0.323339	Mean dependent var	12.36585
Adjusted R-squared	0.320702	S.D. dependent var	7.896350
S.E. of regression	6.508137	Akaike info criterion	6.588627
Sum squared resid	54342.54	Schwarz criterion	6.612653
Log likelihood	-4240.370	Hannan-Quinn criter.	6.597646
F-statistic	122.6149	Durbin-Watson stat	1.897513
Prob(F-statistic)	0.000000		

Ramsey's RESET test in EVeiws



Ramsey's RESET test in EVeiws



Dependent Variable: WAGE
Method: Least Squares
Date: 07/23/16 Time: 12:40
Sample: 1 1289
Included observations: 1289

Variable	Statistic	Prob.
C	071691	0.0000
FEMALE	433184	0.0000
NONWHITE	074139	0.0022
UNION	165626	0.0305
EDUCATION	).79231	0.0000
EXPER	).38205	0.0000

R-squared	0.323339	Mean dependent var	12.36585
Adjusted R-squared	0.320702	S.D. dependent var	7.896350
S.E. of regression	6.508137	Akaike info criterion	6.588627
Sum squared resid	54342.54	Schwarz criterion	6.612653
Log likelihood	-4240.370	Hannan-Quinn criter.	6.597646
F-statistic	122.6149	Durbin-Watson stat	1.897513
Prob(F-statistic)	0.000000		

Ramsey's RESET test in EVeiws

Ramsey RESET Test
Equation: UNTITLED
Specification: WAGE C FEMALE NONWHITE UNION EDUCATION EXPER
Omitted Variables: Powers of fitted values from 2 to 3

	Value	df	Probability
F-statistic	20.12362	(2, 1281)	0.0000
Likelihood ratio	39.87540	2	0.0000

F-test summary:

	Sum of Sq.	df	Mean Squares
Test SSR	1655.358	2	827.6790
Restricted SSR	54342.54	1283	42.35584
Unrestricted SSR	52687.19	1281	41.12973

LR test summary:

	Value	df
Restricted LogL	-4240.370	1283
Unrestricted LogL	-4220.433	1281

Ramsey's RESET test in EVeiws

Unrestricted Test Equation:
Dependent Variable: WAGE
Method: Least Squares
Date: 07/23/16 Time: 12:41
Sample: 1 1289
Included observations: 1289

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.412981	2.453617	1.798561	0.0723
FEMALE	-0.059017	0.797535	-0.073999	0.9410
NONWHITE	-0.195466	0.631646	-0.309454	0.7570
UNION	0.124108	0.564161	0.219987	0.8259
EDUCATION	0.080124	0.302395	0.264966	0.7911
EXPER	0.000969	0.042470	0.022809	0.9818
FITTED^2	0.044738	0.020767	2.154294	0.0314
FITTED^3	-0.000311	0.000601	-0.517110	0.6052

R-squared	0.343951	Mean dependent var	12.36585
Adjusted R-squared	0.340366	S.D. dependent var	7.896350
S.E. of regression	6.413247	Akaike info criterion	6.560795
Sum squared resid	52687.19	Schwarz criterion	6.592830
Log likelihood	-4220.433	Hannan-Quinn criter.	6.572820
F-statistic	95.94255	Durbin-Watson stat	1.894263
Prob(F-statistic)	0.000000		

Ramsey's RESET test: Drawbacks

- although simple to apply, it has two drawbacks

  1. if the model found incorrectly specified, the test does not suggest any specific alternative
  2. does not offer guidance about the number of powered termed

The Lagrange Multiplier LM test

1. from the original model, obtain the estimated residuals, e_i
2. if model is correct, e_i should not be related to the regressors omitted from that model
3. regress e_i on the regressors in the original model and the omitted variables from the original model [this is an auxiliary regression]

The Lagrange Multiplier LM test

4. if large sample, then $n * R^2$ [obtained from the auxiliary regression]
   - follows the chi-square distribution
   - with *df* equal to the number of regressors omitted from the original regression
5. we reject the original [restricted] regression if
   - the computed value exceeds the critical value at the chosen level of significance
   - its p value is sufficiently low

This is, the original model was misspecified

LM test: An Example

Dependent Variable: WAGE
Method: Least Squares
Date: 07/23/16 Time: 12:18
Sample: 1 1289
Included observations: 1289

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-7.183338	1.015788	-7.071691	0.0000
FEMALE	-3.074875	0.364616	-8.433184	0.0000
NONWHITE	-1.565313	0.509188	-3.074139	0.0022
UNION	1.095976	0.506078	2.165626	0.0305
EDUCATION	1.370301	0.065904	20.79231	0.0000
EXPER	0.166607	0.016048	10.38205	0.0000

R-squared	0.323339	Mean dependent var	12.36585
Adjusted R-squared	0.320702	S.D. dependent var	7.896350
S.E. of regression	6.508137	Akaike info criterion	6.588627
Sum squared resid	54342.54	Schwarz criterion	6.612653
Log likelihood	-4240.370	Hannan-Quinn criter.	6.597646
F-statistic	122.6149	Durbin-Watson stat	1.897513
Prob(F-statistic)	0.000000		

LM test: An Example

Dependent Variable: E1
Method: Least Squares
Date: 07/23/16 Time: 13:09
Sample: 1 1289
Included observations: 1289

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-2.017330	1.072115	-1.881636	0.0601
FEMALE	1.640895	0.680797	2.410257	0.0161
NONWHITE	0.083422	0.503577	0.165659	0.8685
UNION	-0.146949	0.501081	-0.293264	0.7694
EDUCATION	-0.051936	0.065801	-0.789287	0.4301
EXPER	0.305367	0.056212	5.432437	0.0000
EXPER^2	-0.006274	0.001224	-5.124559	0.0000
EXPER*FEMALE	-0.084151	0.030848	-2.727939	0.0065

R-squared	0.025089	Mean dependent var	1.15E-15
Adjusted R-squared	0.019761	S.D. dependent var	6.495492
S.E. of regression	6.430992	Akaike info criterion	6.566322
Sum squared resid	52979.16	Schwarz criterion	6.598357
Log likelihood	-4223.994	Hannan-Quinn criter.	6.578346
F-statistic	4.709394	Durbin-Watson stat	1.892702
Prob(F-statistic)	0.000031		

LM test: An Example

- LM test

$$nR^2 = (1289)(0.0251) \approx 32.35 \sim \chi^2_{2,0.05}$$

- the critical value for $\chi^2_{2,.05} = 5.99147$
- the computed value is greater than the critical value
- reject the original model; i.e., it is misspecified

A Chi-Square Table: Values of χ^2_α

df	$\chi^2_{0.10}$	$\chi^2_{0.05}$
1	2.70554	3.84146
2	4.60517	5.99147
3	6.25139	7.81473
4	7.77944	9.48773
5	9.23635	11.0705
6	10.6446	12.5916
7	12.0170	14.0671
8	13.3616	15.5073
9	14.6837	16.9190
10	15.9871	18.3070

Basic Idea

- sometimes researchers add variables hoping that the R^2 value will increase
  - a mistaken belief that the higher the R^2 the better the model
- if variables are not economically meaningful and relevant: overfitting a model

An Example: Wage Determination

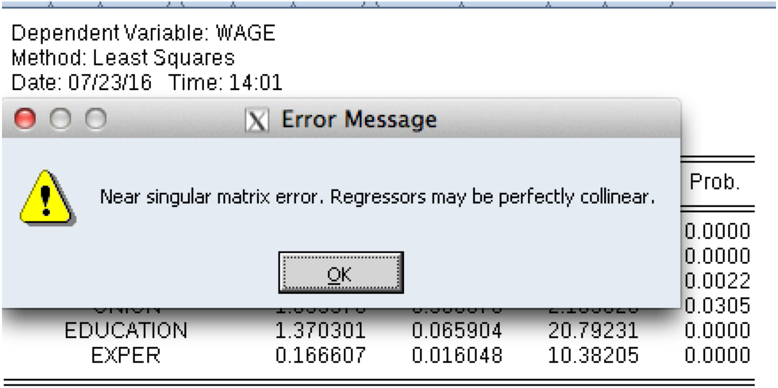
Dependent Variable: WAGE
Method: Least Squares
Date: 07/23/16 Time: 12:18
Sample: 1 1289
Included observations: 1289

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-7.183338	1.015788	-7.071691	0.0000
FEMALE	-3.074875	0.364616	-8.433184	0.0000
NONWHITE	-1.565313	0.509188	-3.074139	0.0022
UNION	1.095976	0.506078	2.165626	0.0305
EDUCATION	1.370301	0.065904	20.79231	0.0000
EXPER	0.166607	0.016048	10.38205	0.0000

R-squared	0.323339	Mean dependent var	12.36585
Adjusted R-squared	0.320702	S.D. dependent var	7.896350
S.E. of regression	6.508137	Akaike info criterion	6.588627
Sum squared resid	54342.54	Schwarz criterion	6.612653
Log likelihood	-4240.370	Hannan-Quinn criter.	6.597646
F-statistic	122.6149	Durbin-Watson stat	1.897513
Prob(F-statistic)	0.000000		

An Example: Wage Determination

- now try adding age of the worker to the model



An Example: Wage Determination

Dependent Variable: EXPER
Method: Least Squares
Date: 07/23/16 Time: 14:06
Sample: 1 1289
Included observations: 1289

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-18.56877	0.269951	-68.78564	0.0000
AGE	0.984808	0.006811	144.5984	0.0000

R-squared	0.942016	Mean dependent var	18.78976
Adjusted R-squared	0.941971	S.D. dependent var	11.66284
S.E. of regression	2.809491	Akaike info criterion	4.905434
Sum squared resid	10158.60	Schwarz criterion	4.913443
Log likelihood	-3159.552	Hannan-Quinn criter.	4.908440
F-statistic	20908.71	Durbin-Watson stat	1.915714
Prob(F-statistic)	0.000000		

Wage Determination: Linear

Dependent Variable: WAGE
Method: Least Squares
Date: 07/23/16 Time: 12:21
Sample: 1 1289
Included observations: 1289

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-9.200668	1.072115	-8.581791	0.0000
FEMALE	-1.433980	0.680797	-2.106326	0.0354
NONWHITE	-1.481891	0.503577	-2.942731	0.0033
UNION	0.949027	0.501081	1.893958	0.0585
EDUCATION	1.318365	0.065801	20.03554	0.0000
EXPER	0.471974	0.056212	8.396344	0.0000
EXPER^2	-0.006274	0.001224	-5.124559	0.0000
EXPER*FEMALE	-0.084151	0.030848	-2.727939	0.0065

R-squared	0.340315	Mean dependent var	12.36585
Adjusted R-squared	0.336711	S.D. dependent var	7.896350
S.E. of regression	6.430992	Akaike info criterion	6.566322
Sum squared resid	52979.16	Schwarz criterion	6.598357
Log likelihood	-4223.994	Hannan-Quinn criter.	6.578346
F-statistic	94.40528	Durbin-Watson stat	1.892702
Prob(F-statistic)	0.000000		

Wage Determination: Log-Linear

Dependent Variable: LNWAGE
Method: Least Squares
Date: 07/23/16 Time: 14:21
Sample: 1 1289
Included observations: 1289

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.732446	0.077613	9.437131	0.0000
FEMALE	-0.148060	0.049285	-3.004179	0.0027
NONWHITE	-0.127302	0.036455	-3.492000	0.0005
UNION	0.168485	0.036275	4.644705	0.0000
EDUCATION	0.094792	0.004764	19.89963	0.0000
EXPER	0.041946	0.004069	10.30778	0.0000
EXPER^2	-0.000637	8.86E-05	-7.187309	0.0000
EXPER*FEMALE	-0.005043	0.002233	-2.258065	0.0241

R-squared	0.373017	Mean dependent var	2.342416
Adjusted R-squared	0.369591	S.D. dependent var	0.586356
S.E. of regression	0.465556	Akaike info criterion	1.315020
Sum squared resid	277.6474	Schwarz criterion	1.347055
Log likelihood	-839.5302	Hannan-Quinn criter.	1.327044
F-statistic	108.8741	Durbin-Watson stat	1.926178
Prob(F-statistic)	0.000000		

Wage Determination Functional Form

1. genr wageaverage = @mean(wage)
2. genr wagenew = wage/wageaverage
3. run both regressions [linear & log-linear] using the new variable and obtain the *RSS* [residual sum of squares] for both models

Wage Determination Functional Form

Dependent Variable: WAGE_NEW
Method: Least Squares
Date: 07/23/16 Time: 14:34
Sample: 1 1289
Included observations: 1289

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.744039	0.086700	-8.581791	0.0000
FEMALE	-0.115963	0.055055	-2.106326	0.0354
NONWHITE	-0.119837	0.040723	-2.942731	0.0033
UNION	0.076746	0.040521	1.893958	0.0585
EDUCATION	0.106613	0.005321	20.03554	0.0000
EXPER	0.038168	0.004546	8.396344	0.0000
EXPER^2	-0.000507	9.90E-05	-5.124559	0.0000
EXPER*FEMALE	-0.006805	0.002495	-2.727939	0.0065

R-squared	0.340315	Mean dependent var	1.000000
Adjusted R-squared	0.336711	S.D. dependent var	0.638561
S.E. of regression	0.520061	Akaike info criterion	1.536444
Sum squared resid	346.4632	Schwarz criterion	1.568479
Log likelihood	-982.2384	Hannan-Quinn criter.	1.548469
F-statistic	94.40528	Durbin-Watson stat	1.892702
Prob(F-statistic)	0.000000		

Wage Determination Functional Form

Dependent Variable: LOG(WAGE_NEW)
Method: Least Squares
Date: 07/23/16 Time: 14:34
Sample: 1 1289
Included observations: 1289

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-1.782493	0.077613	-22.96635	0.0000
FEMALE	-0.148060	0.049285	-3.004179	0.0027
NONWHITE	-0.127302	0.036455	-3.492000	0.0005
UNION	0.168485	0.036275	4.644705	0.0000
EDUCATION	0.094792	0.004764	19.89963	0.0000
EXPER	0.041946	0.004069	10.30778	0.0000
EXPER^2	-0.000637	8.86E-05	-7.187309	0.0000
EXPER*FEMALE	-0.005043	0.002233	-2.258065	0.0241

R-squared	0.373017	Mean dependent var	-0.172523
Adjusted R-squared	0.369591	S.D. dependent var	0.586356
S.E. of regression	0.465556	Akaike info criterion	1.315020
Sum squared resid	277.6474	Schwarz criterion	1.347055
Log likelihood	-839.5302	Hannan-Quinn criter.	1.327044
F-statistic	108.8741	Durbin-Watson stat	1.926178
Prob(F-statistic)	0.000000		

Wage Determination Functional Form

4. compute [larger *RSS* in the numerator]

$$\lambda = \frac{n}{2} ln(\frac{RSS_1}{RSS_2}) \sim \chi^2_{1,\alpha}$$

$$= \frac{1289}{2} ln(\frac{346.4632}{277.6474}) \approx 365.11$$

- the computed value is greater than the critical value [significant]
- then the model with the lower *RSS* value is better
- in this example, log-lin model is superior to the linear model [linear model is misspecified]

Errors of Measurement in the Regressand

- the OLS estimators are still unbiased
- the variances and standard errors of OLS estimators are still unbiased
- but the estimated variances, and ipso facto the standard errors, are larger than in the absence of such errors
- in short, errors of measurement in the regressand do not pose a very serious threat to OLS estimation

Errors of Measurement in the Regressor I

- OLS estimators are biased as well as inconsistent
- errors in a single regressor can lead to biased and inconsistent estimates of the coefficients of the other regressors in the model
- it is not easy to establish the size and direction of bias in the estimated coefficients
- suggested to use instrumental or proxy variables for variables suspected of having measurement errors

Errors of Measurement in the Regressor II

- the proxy variables must satisfy two requirements
  1. highly correlated with the variables for which they are a proxy and
  2. uncorrelated with the usual equation error as well as the measurement error
- more about instrumental variables in chapter 19
- be very careful in collecting the data and make sure that some obvious errors are eliminated