Empirical Economic Research, Part II

Based on the text book by Ramanathan: *Introductory Econometrics*

Robert M. Kunst
robert.kunst@univie.ac.at

University of Vienna
and
Institute for Advanced Studies Vienna

December 7, 2011

Outline

Multiple linear regression: idea

In the multiple linear regression model, a variable depends on several regressors (influence factors):

$$y_t = \beta_1 + \beta_2 X_{t,2} + \ldots + \beta_k X_{t,k} + u_t.$$

Remark: '*multiple*' is not the same as *multivariate*, as the dependent variable is a scalar (not a vector).

Multiple linear regression is the most widely used model of empirical economics.

Explanatory variables are typically not independent

The fundamental problem in interpreting multiple regressions is that the regressors are not mutually independent. For example, is it justified in a consumption function

$$C_t = \beta_1 + \beta_2 YD_t + \beta_3 R_t + \beta_4 UR_t + u_t$$

to interpret $\beta_4 = \frac{\partial C}{\partial UR}$, i.e. as the marginal reaction *ceteris paribus*, if the unemployment rate UR and the aggregate income YD are negatively correlated?

Matrix representation

Stacking the equations for $t = 1, \ldots, n$ vertically yields a matrix representation

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1k} \\ 1 & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix},$$

or in short

$$\vec{y} = \mathbf{X}\vec{\beta} + \vec{u}.$$

Elements of the multiple Regression in matrix notation

- $\overrightarrow{y}$ is an $n \times 1$ vector of observations of the dependent Variable Y.
- $\mathbf{X}$ is a $n \times k$ matrix of observations of the explanatory variables $X_{.1}, \ldots, X_{.k}$. In $x_{t,k}$, the first index denotes the observation, the second index denotes a variable. $X_{.1} \equiv 1$ is convened for the intercept.
- $\overrightarrow{\beta}$ is a $k \times 1$ vector of coefficient parameters to be estimated.
- $\overrightarrow{u}$ is an $n \times 1$ vector of unobserved errors.

Regression as approximate inversion

Assume that $n = k$, and $\mathbf{X}$ is invertible. Then, we can solve

$$\overrightarrow{y} = \mathbf{X}\overrightarrow{\beta}$$

as a system of equations by

$$\overrightarrow{\beta} = \mathbf{X}^{-1}\overrightarrow{y}$$

exactly, the error is 0. This should not be, n is typically much larger than k. Then, we aim for a solution with errors as small as possible, and this is once more OLS.

Almost an inverse matrix

Assume $n > k$, and the $k \times k$ matrix $\mathbf{X}'\mathbf{X}$ is invertible. Then, for the matrix $\mathbf{X}^+ = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ we have

$$\mathbf{X}^+\mathbf{X} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I},$$

but $\mathbf{X}\mathbf{X}^+ \neq \mathbf{I}$. O.c.s. that the estimator

$$\vec{\hat{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\vec{y}$$

is just the OLS estimator for the multiple linear regression. Proof by minimization and k–dimensional differential. A simple representation for the estimator without matrices does not exist. (From now on, the little arrows for vectors will be suppressed.)

Multicollinearity technically

The OLS estimator can only be calculated, if $\mathbf{X}'\mathbf{X}$ is invertible (non-singular). This requires that the rank of $\mathbf{X}$ is k, i.e. that the columns of $\mathbf{X}$ are linear independent. No column of $\mathbf{X}$ can be expressed as a linear combination of the others.

Otherwise, $\mathbf{X}$ will have rank $< k$. This situation is called *multicollinearity*.

The interpretation of multicollinearity

'Exact' multicollinearity is the consequence of a conceptual mistake. If the regressors are linear dependent, the result becomes arbitrary.

Example: Consider the consumption function

$$C_t = \beta_1 + \beta_2 X_{t2} + \beta_3 X_{t3} + \beta_4 X_{t4} + u_t,$$

with X_2 the disposable income, X_3 the aggregate disposable wage income, X_4 the non-wage income. Obviously, no reasonable result can be obtained here.

Dummy variable trap

Quarterly dummy variables, as often used with seasonal data, form together with the regression constant X_1 5 linear dependent columns in $\mathbf{X}$: multicollinearity.

Remedy: remove one dummy variable (technically preferable) or estimate homogeneously. Both methods yield the same residuals and the same estimator for the remaining coefficients.

The same effect occurs for: dummies male and female; dummies Democrat and Republican president etc.

Near-multicollinearity

If several explanatory variables are similar with regard to their content or definition, OLS can be calculated but the estimation will become very uncertain (large variance). **Example**: two interest rates in an investment function. Recommended procedure: careful elimination of single suspicious regressors, comparison of several variants etc.

Incorrect usage of the word: mere correlation among regressors is **not** multicollinearity. Explanatory variables are typically not independent.

New assumptions in the multiple model

Most assumptions for the simple model can likewise be applied to the multiple model. The three assumptions to be revised are # 1, # 2,# 7.

Assumption 1 is to be replaced by Assumption $\mathbf{1_M}$:

$$y_t = \beta_1 + \beta_2 X_{t,2} + \ldots + \beta_k X_{t,k} + u_t.$$

Assumption 2 is to be replaced by Assumption $\mathbf{2_M}$: The matrix $\mathbf{X}$ has rank k. In other words, there is no multicollinearity, and no explanatory variable can be represented as a linear combination of other explanatory variables.

Unbiasedness and consistency

Under Assumptions 1_M, 2_M, 3,4 will the OLS estimator in the multiple model

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

be unbiased and, under some additional technical auxiliary assumptions, also consistent for the vector $\beta = (\beta_1, \ldots, \beta_k)'$. Unbiasedness and convergence are defined componentwise.

Sketch of proof for expectation:

$$\begin{aligned} \mathrm{E}\hat{\beta} &= \mathrm{E}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathrm{E}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u}) \\ &= \beta + 0 \end{aligned}$$

Linear efficiency in the multiple model

Under Assumptions 1_M, 2_M, 3–6 will the OLS estimator in the multiple model be the efficient linear unbiased estimator. It has the smallest variance among these estimators: **BLUE** (*best linear unbiased estimator*).

If even Assumption 8 holds, then the OLS estimator is even efficient among all unbiased estimators.

Assumption 7 in the multiple model

Assumption 7 from the simple model is to be replaced by Assumption 7_M: The number of observations n is strictly larger than the number of estimated coefficients k.

The case $n < k$ is excluded by Assumption 2_M. For $n = k$ will the regression fit perfectly, all residuals are 0, and no statement on variances is possible.

A meaningful regression requires that this assumption holds 'very' strictly, n should be considerably larger than k. From 8 observations, it is not possible to estimate 6 or 7 coefficients reliably.

Variance of the OLS estimator

Under Assumptions 1–6, the variance of the OLS estimator can be shown to be

$$\mathrm{var}\hat{\beta} = \sigma_u^2(\mathbf{X}'\mathbf{X})^{-1}.$$

- $\hat{\beta}$ is a vector, the variance is a matrix, a 'variance-covariance matrix', whose diagonal carries the variances of the coefficient estimators $\sigma^2_{\beta,j}$;
- a simple formula without matrix notation does not exist;
- Proof not difficult.

Details of the variance matrix

$$\mathrm{var}\begin{pmatrix} \hat\beta_1 \\ \hat\beta_2 \\ \vdots \\ \hat\beta_k \end{pmatrix} = \begin{pmatrix} \mathrm{var}(\hat\beta_1) & \mathrm{cov}(\hat\beta_1,\hat\beta_2) & \cdots & \mathrm{cov}(\hat\beta_1,\hat\beta_k) \\ \mathrm{cov}(\hat\beta_1,\hat\beta_2) & \mathrm{var}(\hat\beta_2) & & \mathrm{cov}(\hat\beta_2,\hat\beta_k) \\ \vdots & & \ddots & \vdots \\ \mathrm{cov}(\hat\beta_1,\hat\beta_k) & \mathrm{cov}(\hat\beta_2,\hat\beta_k) & \cdots & \mathrm{var}(\hat\beta_k) \end{pmatrix}$$

$$= \sigma_u^2 \begin{pmatrix} n & \sum x_{t2} & \cdots & \sum x_{tk} \\ \sum x_{t2} & \sum x_{t2}^2 & & \sum x_{t2}x_{tk} \\ \vdots & & \ddots & \vdots \\ \sum x_{tk} & \sum x_{t2}x_{tk} & \cdots & \sum x_{tk}^2 \end{pmatrix}^{-1}$$

Estimating the error variance

The variance of the OLS estimator depends on the unknown σ_u^2, which is estimated by

$$\hat{\sigma}_u^2 = \frac{1}{n-k} \sum_{t=1}^{n} \hat{u}_t^2 .$$

The estimator is unbiased under Assumptions $1_M, 2_M$,3–6,7_M.

- The denominator $n - k$ corrects for the degrees of freedom lost in OLS estimation;
- the square root $\hat{\sigma}_u$ is an estimator for the standard deviation with a tolerable bias.

R^2 in the multiple model

In analogy to simple regression, an indicator of determination R^2 can be defined via

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{t=1}^{n} \hat{u}_t^2}{\sum_{t=1}^{n} (y_t - \bar{y})^2}.$$

This R^2 is the maximum empirical correlation between Y and linear combinations of $X_2, \ldots, X_k$.

Problem: R^2 increases (or remains the same), if a further explanatory variable is added to the regressors. With growing k will R^2 approach 1. R^2 does not enable a reliable comparison among regressions with different k.

Corrected R^2 in the multiple model

In the multiple model, the corrected (or adjusted) R^2 according to WHERRY and THEIL is defined as

$$\bar{R}^2 = 1 - \frac{RSS/(n-k)}{TSS/(n-1)} = 1 - \frac{n-1}{n-k}(1 - R^2).$$

This $\bar{R}^2$ estimated the maximum correlation between random variables Y and linear combinations of random variables $X_2, \ldots, X_k$ with a smaller bias than R^2.

- Because of Assumption 4, our regressors are not stochastic, hence we do not really estimate anything;
- As a descriptive coefficient of determination, $\bar{R}^2$ is hardly preferable to R^2.

Properties of $\bar{R}^2$ in the multiple model

- $\bar{R}^2$ for $k \uparrow n$ does not approach 1;
- $\bar{R}^2$ may even fall after adding regressors, whereas it still increases ‘too easily’;
- $\bar{R}^2$ can even become negative: this indicates a very bad regression.

Homogeneous variants of R^2

With homogeneous regressions (without constant, $X_1 \neq 1$), many regression programs provide a homogeneous R^2

$$R_0^2 = 1 - \frac{\sum_{t=1}^{n} \hat{u}_t^2}{\sum_{t=1}^{n} y_t^2}.$$

Comparisons to corresponding inhomogeneous regressions will be unreliable: homogeneous regressions should be used sparingly.

Information criteria: the idea

Specification searches are one of the main issues in empirical economics: the researcher aims for the best explanation of a given dependent variable Y. Many regressions with different regressors and different k are 'tried out'. What should be the criterion for the best variant in the comparison?

R^2 and $\bar{R}^2$ are not adequate. Information criteria are adequate. Each variant obtains a statistic (a number) that falls with better determination and increases with k. The smallest number then indicates the optimum.

A simple information criterion: estimated error variance

The unbiased estimate of the error variance

$$\hat{\sigma}_u^2 = \frac{1}{n-k} \sum_{t=1}^{n} \hat{u}_t^2$$

is a simple information criterion. The factor $\sum \hat{u}_t^2$ falls if more regressors are added and k increases, the factor $\frac{1}{n-k}$ rises.

Minimizing this criterion means to maximize $\bar{R}^2$, which is not good: the 'penalty term' for the loss in degrees of freedom $\frac{1}{n-k}$ is too weak.

The criterion AIC

The criterion AIC, which was developed by AKAIKE, uses a penalty term that increases faster. RAMANATHAN uses the antilog of AIC, which implies equivalent properties regarding model selection:

$$AIC = \frac{RSS}{n} \exp\left(\frac{2k}{n}\right).$$

Under certain conditions one can show that the model that is determined by minimization of AIC has good properties. For large n, the forecast for y implied by the minimizing model will be optimal.

The Schwarz criterion

The modified BIC developed by Gideon Schwarz (often the acronym is interpreted as *Bayes Information Criterion*), in the antilog version of Ramanathan

$$SBIC = \frac{RSS}{n} n^{k/n},$$

has a more severe penalty term than AIC.

Under certain conditions one can show that the model determined by minimizing SBIC will, for $n \to \infty$, contain exactly those regressors, whose influence on Y is not zero ('asymptotically true model'). For small n, many econometricians may prefer AIC.

Misunderstandings about information criteria

- Definitions of information criteria have not been standardized. Monotonic transformations of AIC or SBIC choose the same model. Different software programs and books: different values.
- Values of information criteria make sense in a comparison only, in isolation they are meaningless. Advantage of $\bar{R}^2$: interpretable.
- Information criteria do not contradict hypothesis tests. Under certain conditions will the decision according to criteria be equivalent to a decision by a hypothesis test at a specific significance level (AIC $>$ 10%, SBIC $\downarrow$ 0% for $n \to \infty$).
- Information criteria (and also $\bar{R}^2$) can compare arbitrary models, hypothesis tests can compare 'nested' cases only: AIC and SBIC are stronger tools.

Tests as tools of Specification search

For the specification search, the following tools are available:

1. Restriction tests, in order to decide whether specific regressors should be in the regression or not: simple versus general models;
2. Information criteria, to compare arbitrary specifications;
3. Specification tests, in order to test whether specific model assumptions hold.

The t–Test for $H_0 : \beta_j = 0$

The t–statistic for the coefficients β_j

$$t_{\beta_j} = \frac{\hat{\beta}_j}{\hat{\sigma}_{\beta_j}}$$

is under H_0 t–distributed with $n - k$ degrees of freedom, if Assumptions 1–8 (with M) hold. $\hat{\sigma}_{\beta_j}$ has been introduced already. Without Assumption 8, the normal distribution is to be used for large n.

Rule of thumb: 't is significant, if > 2' in absolute value.

The F–test for $H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0$

All coefficients except the intercept are tested jointly. The alternative is that at least one β_j is not 0. The F-total statistic

$$F_c = \frac{ESS/(k-1)}{RSS/(n-k)} = \frac{R^2}{1-R^2}\frac{n-k}{k-1}$$

is under H_0 F–distributed with $(k-1, n-k)$ degrees of freedom, if Assumptions 1–8 (with M) hold.

Test 'should' reject: if H_0 is true, there does not exist any meaningful regression model, no regressor is useful.

F–Test for several coefficients, but not all

Presume that we wish to test $H_0 : \beta_2 = \beta_3 = \ldots = \beta_l = 0$, i.e. $l - 1$ variables. The *unrestricted* model (U) under H_A is

$$y_t = \beta_1 + \beta_2 x_{t2} + \ldots + \beta_k x_{tk} + u_t,$$

the *restricted* model (R) under H_0 is

$$y_t = \beta_1 + \beta_{l+1} x_{t,l+1} + \ldots + \beta_k x_{t,k} + u_t.$$

Even for this situation, an F–Test is appropriate.

The restricted model has $DF_R = n - k + l - 1$ degrees of freedom, the unrestricted model has $DF_U = n - k$ degrees of freedom, the difference is $l - 1$.

The general F–statistic

The F–statistic for both models (U) and (R)

$$F_c = \frac{(RSS_R - RSS_U)/(DF_R - DF_U)}{RSS_U/DF_U}$$

is calculated from the residual sums of squares for both models. It is under the H_0–model (R) F–distributed with $(l-1, n-k)$ degrees of freedom, if Assumptions 1–8 (with M) hold.

Such F–statistics are not supplied automatically, only on request: all potential combinations would be too many.

Tools for specification search: comparison

- R^2 and $\bar{R}^2$ are not well suited for a comparison of specifications: R^2 always increases after adding regressors, $\bar{R}^2$ much too often;
- Information criteria can compare all specifications: minimum is 'statistical' optimum, need not coincide with economic optimum;
- $F-$ and $t-$Tests can compare chains of 'nested' models only: problems with significance levels, 'general-to-specific' elimination or 'specific-to-general' inclusion, often equivalence with information criteria.

Specification tests: Idea

- *Restriction tests* assist in the decision between specifications under (not tested) assumptions: null hypothesis is restricted model;
- *Specification tests* examine the assumptions of the regression model: null hypothesis is validity of the assumption.

Rejection in specification tests is often undesired: warning sign, quality control.

Which assumptions are typically tested?

1. Tests for correctness of the functional form (linear versus nonlinear) exist;
2. Multicollinearity is no stochastic property and can be checked without statistical tests;
3. Assumption 3 is an auxiliary assumption and cannot be tested;
4. Tests for exogeneity, causality can be interpreted as checks of Assumption 4;
5. Tests for heteroskedasticity exist;
6. Tests for autocorrelation are contained in standard printouts: Durbin-Watson test etc.;
7. Assumption 7 can be checked without tests;
8. Tests for normality of errors exist: Jarque-Bera test etc.

Durbin-Watson test: idea

The DW–Test tests the null hypothesis (=Assumption 6) that the errors of the regression are uncorrelated across t (not *autocorrelated*).

The alternative is that the errors u_t follow an autoregressive model of order one (AR(1)):

$$u_t = \rho u_{t-1} + \varepsilon_t$$

Widespread mistake: This test neither looks for autocorrelation in observed variables nor in the residuals $\hat{u}_t$.

Durbin-Watson test statistic

Naively, one might consider

$$\operatorname{corr}(\hat{u}_t, \hat{u}_{t-1}) = \hat{\rho} = \frac{\sum_{t=2}^{n} \hat{u}_t \hat{u}_{t-1}}{\sum_{t=1}^{n} \hat{u}_t^2}.$$

The actual Durbin-Watson statistic is similar:

$$d = \frac{\sum_{t=2}^{n} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{n} \hat{u}_t^2}$$

The correspondence $d \approx 2(1 - \hat{\rho})$ holds approximately.

Properties of the DW–statistic

- $0 \leq d \leq 4$;
- $d \in (0, 2)$ for $\rho > 0$, positive autocorrelation (occurs often);
- $d \in (2, 4)$ for $\rho < 0$, negative autocorrelation (occurs rarely);
- $d \approx 2$ if $\rho = 0$, i.e. if H_0 holds;
- The distribution of d under H_0 (and Assumptions 1–8) is no standard distribution.

Distribution of d under H_0

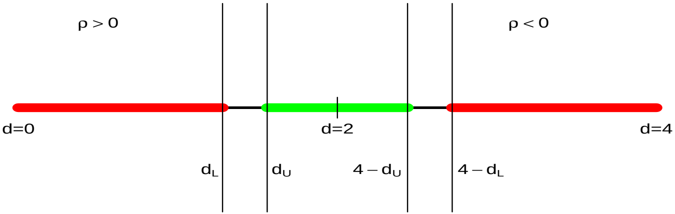
The distribution of d under $H_0 : \rho = 0$ depends on the regressor matrix $\mathbf{X}$, general significance points cannot be tabulated.

2 solutions:

- Computer determines exact significance points for a given regression problem;
- Maxima and minima of the significance points are tabulated.

d_U and d_L

Under Assumptions 1–8 and for given n and $k' = k - 1$, tables of maximal (d_U) and minimal significance points d_L are available. Traditional usage:



Between d_L and d_U, no decision is taken.

Durbin-Watson test: pro and contra

+ The DW test is under Assumptions 1–8 (normal regression) an exact test;
+ The DW test is also sensitive to violations of assumptions other than Assumption 6: general test for mis-specification;
- The distribution of the DW–statistic is uncontrollable in dynamic regressions such as $y_t = \beta_1 + \beta_2 y_{t-1} + ... + u_t$, the test must not be used in such cases (violation of Assumption 4);
- The alternative is very special, other forms of autocorrelation are not captured.

Lagrange-Multiplier tests: the idea

The LM principle is one of the two important approximations to the LR test (the other one is the Wald test). It can be applied to many testing problems.

Technically, the model is estimated under H_0, and the derivative of the likelihood 'in the direction of the alternative' is investigated: what do we gain if the model is made more general in this direction?

Under certain conditions is the LM statistic for large n under H_0 $\chi^2(m)$–distributed, with m equal to the number of restrictions that separate H_0 and H_A.

Examples for LM tests

- The Jarque-Bera test for normal distribution (Assumption 8) in the errors and also in variables is a LM–Test, the statistic is asymptotically $\chi^2(2)$ distributed (skewness, kurtosis);
- The Breusch-Pagan test for omitted explanatory variables is an LM test that is used instead of the F–Test: asymptotically $\chi^2(m)$ distributed, with m denoting the number of considered additional regressors;
- The portmanteau Q statistic summarizes autocorrelation in the errors generally, an approximative LM–statistic, with uncorrelated errors asymptotically χ^2 distributed with uncertain degrees of freedom;
- In contrast to the DW test, the LM test for autocorrelation can even be applied to dynamic models.

LM–Test for autocorrelation

This test is an approximative LM test that uses the form of nR^2. Its null hypothesis is the regression model

$$y_t = \beta_1 + \beta_2 x_{t,2} + \ldots + \beta_k x_{t,k} + u_t$$

with Assumption 6. The alternative is an autoregressive model

$$u_t = \rho_1 u_{t-1} + \ldots + \rho_m u_{t-m} + \varepsilon_t,$$

with uncorrelated ε_t.

In the construction of the test statistic, u_t are replaced by residuals $\hat{u}_t$, but differently from DW.

Main regression and auxiliary regression

1. Estimate the model

$$y_t = \beta_1 + \beta_2 x_{t,2} + \ldots + \beta_k x_{t,k} + u_t$$

   by OLS: this yields residuals $\hat{u}_t$;
2. Regress $\hat{u}_t$ on lags:

$$\hat{u}_t = \gamma_1 + \gamma_2 x_{t,2} + \ldots + \gamma_k x_{t,k} + \phi_1 \hat{u}_{t-1} + \ldots + \phi_m \hat{u}_{t-m} + e_t,$$

   with error term e_t: this auxiliary regression has an R^2;
3. the LM–statistic $(n-m)R^2$ is under H_0 $\chi^2(m)$ distributed, if n is large.

Remark: The regressors $x_{t,2}$ etc. in the auxiliary regression are important for theoretical reasons, $\hat{\gamma}_j, j = 1, \ldots, k$ should be small.