ECON 4160, Autumn term 2014. Lecture 2

Ragnar Nymoen

University of Oslo

25 August 2014

References to Lecture 2

- HN: CH: 6 (matrix algebra simple regression),7,8 (matrix algebra multivariate regression)
- Lecture note 1 on the web-page.
- DM:Ch 1-4.5, 5.1-5.2.

Regression models and other model equations I

- Hendry and Nilsen (HN), but also Davidson and MacKinnon (DM), start with the joint probability function (pdf) for the observable random variables.
- Extending the joint distribution of the data from Lecture 1, we can think of

  $$f(X_0, X_1, \ldots, X_k)$$

  as the pdf for the $k+1$ random variables $(X_{0i}, X_{1i}, \ldots, X_{ki})$.
- When we choose to use a *regression model equation*, we specify one random variable as a regressand and the other as regressors.

Regression models and other model equations II

- Let

  $$Y = X_0$$

  define the regressand.

- We can always write the joint pdf as the product of a conditional pdf and a marginal pdf:

  $$f(Y, X_1, \ldots, X_k) = f(Y \mid X_1, \ldots, X_k) \cdot f(X_1, \ldots, X_k) \quad (1)$$

  - Note that $f(X_1, \ldots, X_k)$ is a joint pdf in its own right, but it is marginal **relative to** the full pdf on the left hand side.

Regression models and other model equations III

- From $f(Y \mid X_1, \ldots, X_k)$ we can always construct the **conditional expectation function**:

$$E(Y \mid X_1, \ldots, X_k)$$

and the **disturbance**

$$\varepsilon = Y - E(Y \mid X_1, \ldots, X_k).$$

- For a realization of the X-variables, the conditional expectation $E(Y \mid x_1, \ldots, x_k)$ is deterministic. But we can consider the expectation for any realization of X, and by the **Law of iterated expectations** we get:

$$E(E(\varepsilon \mid X_1, \ldots, X_k)) = E(\varepsilon)$$

$$E([Y - E(Y \mid X_1, \ldots, X_k)] \mid X_1, \ldots, X_k) = 0 \Longrightarrow E(\varepsilon) = 0$$

Regression models and other model equations IV

- For the population regresson:

$$Y = E(Y \mid X_1, \ldots, X_k) + \varepsilon$$

  we want to estimate the parameters of $E(Y \mid X_1, \ldots, X_k)$.

- To be relevant, the parameters of $E(Y \mid X_1, \ldots, X_k)$ should correspond to the *parameters of interest* for our research. Examples: The average response in Y to a change in one X_j. The best prediction of Y given $x_1, \ldots, x_k$.
- An altenative to regression modelling is instead to put the joint pdf $f(Y, X_1, \ldots, X_k)$ on model form. We call this **system modelling.**

Regression models and other model equations V

- The distinction between regression model equations and other model equations cas be subtle: For example, we often focus on a single equation in the joint pdf.
- This is one (of two ways, which is the other?) that you are introduced to IV estimation in your introductory course. The model equation may "look like" a linear regression model:

  $$Y = \gamma_1 + \gamma_2 X_2 + \gamma_3 X_3 + \epsilon \tag{2}$$

  but because $E(\epsilon \mid X_1, X_2) \neq 0$, it is **not** a regression equation but a **structural equation** which is part of a model representation of the the joint pdf.
- Terminology like "regression with endogenous variables" should be avoided.

- The assumptions of the statistical model are identical to Ch. 5 in HN (end of slide set 1), but we allow for k regressors
- $k = 2$ in Ch 5 in HN and $k = 1$ in Ch 3 of HN because their k includes the constant.
- Of course in other books k does not count the constant term.
- The likelihood function is constructed from the **conditional pdf**: $f(Y \mid X_1, \dots, X_k)$.
- Assume **identical distributions** of n **independent** sets of variables (relevant for cross-section data)

$$f(Y_1, \dots Y_n \mid X_{11}, \dots, X_{kn}) = \prod_{i=1}^{n} f(Y_i \mid X_{1i}, \dots, X_{ki})$$

- Assume that each of the n conditonal pdfs are **normal**, the likelihood is then:

$$L = \prod_{i=1}^{n} f(Y_i \mid X_{1i}, \ldots, X_{ki}) =$$

$$(2\pi\sigma^2)^{-n/2} \exp\left\{ \frac{-1}{2\pi\sigma^2} \sum_{i=1}^{n} (Y_i - \textstyle\sum_{j=1}^{k} \beta_j X_{ji})^2 \right\}$$

and the corresponding log-likelihood function

$$l_{Y_1,\ldots,Y_n|\mathbf{X}} = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\pi\sigma^2}\sum_{i=1}^{n}(Y_i - \textstyle\sum_{j=1}^{k}\beta_j X_{ji})^2 \tag{3}$$

The log-likelihood function of Ch 5 in HN is obtained by setting $k = 2$. (In Ch 3 by setting $k = 1$, and setting $X_{1i} = 1$ for all i)

- As in that case we can find the MLEs for $\beta_1,\dots,\beta_k$ by applying the OLS principle to the second part of the expression: Therefeore OLS estimators of the β's are MLE, and vise versa.
- Next, from the concentrated log-likelihood function for σ^2, we find the MLE of the scale parameter as:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

where the residuals are from the k-variable model of course.

The regression model in matrix notation I

Let $\mathbf{X}$ be a $n \times k$ matrix with the regressors of the model

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{4}$$

where $\mathbf{y}$ is $n \times 1$and $\boldsymbol{\varepsilon}$ is the $n \times 1$ vector with disturbances and the parameter vector $\boldsymbol{\beta}$ is $k \times 1$.
Notation convention: Use lowercase bold for data vectors.
Uppercase bold for data matrices.

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1k} \\ X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & & \vdots \\ X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The regression model in matrix notation II

if we let $\mathbf{X}_i$ denote the i^{th} row in $\mathbf{X}$. ($1 \times k$ matrix), a typical row in equation (4) is:

$$Y_i = \mathbf{X}_i \boldsymbol{\beta} + \varepsilon_i = \sum_{j=1}^{k} X_{ij}\beta_j + \varepsilon_i,\ i = 1, 2, \ldots, n \tag{5}$$

Unless both the regressand and all the regressors are measured as deviations from their means, there is an intercept in the model. When we need to make this explicit, we can rewrite $\mathbf{X}$ as the *partitioned* matrix:

$$\mathbf{X} = \left[\begin{array}{ccc} \boldsymbol{\iota} & \vdots & \mathbf{X}_2 \end{array}\right]$$

The regression model in matrix notation III

where

$$\boldsymbol{\iota} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}_{n \times 1}$$

$$\mathbf{X}_2 = \begin{bmatrix} X_{12} & \dots & X_{1k} \\ X_{22} & \dots & X_{2k} \\ \vdots & & \vdots \\ X_{n2} & \dots & X_{nk} \end{bmatrix}_{n \times (k-1)}$$

ML estimator I

- By solving Exercise B to the first seminar, you will show that both the Method-of-Moments (MM) and the Ordinary Least Squares (OLS) principle gives the estimator

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{6}$$

  for $\boldsymbol{\beta}$.
- Here, $(\mathbf{X}'\mathbf{X})^{-1}$ is the inverse of the $\mathbf{X}'\mathbf{X}$ matrix with (uncentered) moments between the regressors.
- For the inverse to exist, $rank(\mathbf{X}'\mathbf{X}) = k$, (full rank). This is the generalization of the "absence of perfect multicollinearity" condition.

ML estimator II

- DM uses $\mathbf{X}^{\mathrm{T}}$ as symbol for the **transpose**. HN use the more common $'$ notation, which also avoids confusion with T for the number of observations of time series data.

$$l_{Y_1,\ldots,Y_n|\mathbf{x}} = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\pi\sigma^2}(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)$$

- $\hat{\boldsymbol{\beta}}$ in (6) is minimizes $(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$ meaning that is also the MLE of $\boldsymbol{\beta}$.
- By solving the first Exercise B to the first seminar, you can show that $\hat{\boldsymbol{\beta}}' = \begin{pmatrix} \hat{\beta}_1 & \vdots & \hat{\boldsymbol{\beta}}_2' \end{pmatrix}$ and

$$\hat{\boldsymbol{\beta}}_2 = \left[(\mathbf{X}_2-\bar{\mathbf{X}}_2)'(\mathbf{X}_2-\bar{\mathbf{X}}_2)\right]^{-1}(\mathbf{X}_2-\bar{\mathbf{X}}_2)'\mathbf{y} \qquad (7)$$

In $\bar{\mathbf{X}}_2$, the typical row is $\boldsymbol{\iota}\bar{X}_i$, $i = 2,\ldots,k$.

ML estimator III

- (7) is the generalization of our old friend the "x-deviation from mean" form of OLS estimators that a course in elementary econometrics study in detail for the case of $k = 1$ and $k = 2$! With $k = 2$, Making the two regressors orthogonal to the reparameterized contant term.
- In § 7.2.2. in HN, they take the approach a step further and orthonalize the two (non-contant) regressors also with respect to each other.
- Of course, this can only be achieved by re-parameterizing not only the constant, but also one of the regression coefficientes, see p. 101 in HN for example
- This is fine for understanding how the maximum may be achived in steps also in the k variable case.

ML estimator IV

- But in practice, it is only helpfull if the parameters of interest are the re-parameterized coefficients.

Properties of OLS estimators I

- (6) and (7) are "only" matrix formulations of the OLS estimators for multiple regression that we know from before, it is clear that all the properties that we know from an introductory course still hold.
- Specifically

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

reminds us that the conditional expectation $E(\hat{\beta} \mid \mathbf{X})$ and variance $Var(\hat{\beta} \mid \mathbf{X})$ depend on the assumption about the disturbances in ε.

Properties of OLS estimators II

- With reference to the exogenity assumption of the model specifcation (in HN) it follows that $E(\hat{\boldsymbol{\beta}} \mid \mathbf{X}) = \boldsymbol{\beta}$
- and $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ from the law of iterated expectations.
- Ch 3.1-3.5 in DM is a good exposition for reviewing of the Gauss-Markov/BLUE theorem and other results for the classical regression model.l

Two important matrices I

- From Ch 2.3 in DM we highlight two important matrices in regression theory
- The “residual maker”

$$\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{8}$$

plays a central role in many derivations.

Two important matrices II

- The name stems for the fact that

$$\begin{aligned}\mathbf{My} &= \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{y} - \mathbf{X}\hat{\beta} \equiv \hat{\varepsilon}\end{aligned}$$

  The following properties are worth noting:

$$\begin{aligned}\mathbf{M} &= \mathbf{M}', \text{ symmetric matrix} \\ \mathbf{M}^2 &= \mathbf{M}, \text{ idempotent matrix} \\ \mathbf{MX} &= \mathbf{0}, \text{ regression of } X \text{ on } X \text{ gives perfect fit}\end{aligned}$$

Two important matrices III

- The prediction matrix

  $$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{9}$$

  gives

  $$\hat{\mathbf{y}} = \mathbf{P}\mathbf{y} = \mathbf{X}\hat{\beta}$$

  It is also symmetric and idempotent.
- **M** and **P** are orthogonal:

  $$\mathbf{MP} = \mathbf{PM} = \mathbf{0}$$

  DM say that they **annihiliate** each other.

Two important matrices IV

M and **P** are **complementary projections**

$$\mathbf{M} + \mathbf{P} = \mathbf{I} \tag{10}$$

which gives

$$\mathbf{y} = \hat{\mathbf{y}} + \hat{\varepsilon} = \mathbf{P}\mathbf{y} + \mathbf{M}\mathbf{y}$$

TSS, ESS and all that I

From

$$\mathbf{y} = \hat{\mathbf{y}} + \hat{\varepsilon} = \mathbf{P}\mathbf{y} + \mathbf{M}\mathbf{y}$$

we get

$$\begin{aligned}\mathbf{y}'\mathbf{y} &= (\mathbf{y}'\mathbf{P} + \mathbf{y}'\mathbf{M})(\mathbf{P}\mathbf{y} + \mathbf{M}\mathbf{y}) \\ &= \hat{\mathbf{y}}'\hat{\mathbf{y}} + \hat{\varepsilon}'\hat{\varepsilon}\end{aligned}$$

Written out, this is:

$$\underbrace{\sum_{i=1}^{n} Y_i^2}_{TSS} = \underbrace{\sum_{i=1}^{n} \hat{Y}_i^2}_{ESS} + \underbrace{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}_{RSS} \qquad (11)$$

TSS, ESS and all that II

You may be more used to write this famous decomposition as:

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{TSS} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \overline{\hat{Y}})^2}_{ESS} + \underbrace{\sum_{i=1}^{n}\hat{\varepsilon}_i^2}_{RSS} \tag{12}$$

There is no conflict here, since

$$\mathbf{X}'\hat{\varepsilon} = \mathbf{X}'(\boldsymbol{y} - \mathbf{X}\hat{\beta}) \underset{\substack{MM\\OLS}}{=} \mathbf{0},$$

and, when there is an intercept in the regression:

$$\mathbf{X} = \left[\begin{array}{ccc} \boldsymbol{\iota} & \vdots & \mathbf{X}_2 \end{array}\right]$$

TSS, ESS and all that III

this gives

$$\boldsymbol{\iota}'\hat{\boldsymbol{\varepsilon}} = \sum_{i=1}^{n} \hat{\varepsilon}_i = 0 \tag{13}$$

in the first row of $\mathbf{X}'\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$, and therefore:

$$\bar{Y} = \overline{\hat{Y}}. \tag{14}$$

Using this, you can show that (12) can be written as (11).

TSS, ESS and all that IV

- Memo: Alternative ways of writing *RSS*:

$$\hat{\varepsilon}'\hat{\varepsilon} = (\mathbf{y}'\mathbf{M}')\mathbf{M}\mathbf{y} = \mathbf{y}'\mathbf{M}\mathbf{y} = \mathbf{y}'\hat{\varepsilon} = \hat{\varepsilon}'\mathbf{y} \tag{15}$$

$$\hat{\varepsilon}'\hat{\varepsilon} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{P}'\mathbf{P}\mathbf{y} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{y} \tag{16}$$

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

TSS, ESS and all that V

- The coefficient of determination (R^2) is defined with reference to (12):

$$R^2 = 1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \\ = 1 - \frac{\hat{\varepsilon}'\hat{\varepsilon}}{(\mathbf{M}_{\iota}\mathbf{y})'(\mathbf{M}_{\iota}\mathbf{y})}$$

$$\mathbf{M}_{\iota} = \mathbf{I} - \frac{1}{n}\boldsymbol{\iota}\boldsymbol{\iota}'$$

where we have used the the idempotent centring matrix $\mathbf{M}_{\iota}$, see Lecture note 1 for example.

LR-tests and correlation coefficients I

- Consider the $k = 3$. As a direct extension of §5.5.2 in HN, the LR test of H_0: $\beta_2 = \beta_3 = 0$ in the classical regression case is shown to be

$$LR_{\beta_2=\beta_3=0} = -n\ln(1 - R^2) \overset{D}{\approx} \chi^2(2) \tag{17}$$

- Of course, if the exact normality assumption of the conditional distribution holds, the finite sample test can be used:

$$F_{\beta_2=\beta_3=0} = \frac{R^2}{1-R^2}\frac{n-3}{2} \overset{D}{=} F_{\beta_2=\beta_3=0}(2, n-3)$$

LR-tests and correlation coefficients II

- For a single hypothesis: H_0: $\beta_3 = 0$ the expression of the LR-statistic is:

$$LR_{\beta_3=0} = -n \ln(1 - r^2_{Y,X_3|X_2}) \overset{D}{\approx} \chi^2(1) \qquad (18)$$

  where $r_{Y,X_3|X_2}$ is the **partial correlation coefficient** between Y and X_3 (HN use notation $r_{y,3\cdot1,2}$)

- $r_{Y,X_3|X_2}$ is the correlation coefficient between the residuals from two regressions: Regressing Y and a constant and X_2, and regressing X_3 on X_2 and a constant, see Frisch-Waugh Theorem part of Lecture Note 1.

LR-tests and correlation coefficients III

- Finally, if $\beta_3 = 0$ is true, then $\beta_2 = \beta_3 = 0$ can be tested within the $k = 2$ model as

$$LR_{\beta_2=0|\beta_3=0} = -n \ln(1 - r^2_{Y,X_2}) \overset{D}{\approx} \chi^2(1) \qquad (19)$$

where $r^2_{Y,X_2} = R^2$ in the $k = 2$ model.

Multiple testing I

- The three LR-statistics above are related by

$$LR_{\beta_2=\beta_3=0} = LR_{\beta_3=0} + LR_{\beta_2=0|\beta_3=0} \tag{20}$$

- Within the $k = 3$ model, the two statistics on the right hand side are independent. This means that if we test

$$H_0 : \beta_2 = \beta_3 = 0$$

by *two tests*, the overall *Type-I* error probability is

$$\begin{aligned} P(LR_{\beta_3=0} < c_a) \cdot P(LR_{\beta_2=0|\beta_3=0} < c_\alpha) &\approx (1-\alpha)^2 = 1 - 2\alpha + \alpha^2 \\ &\approx 1 - 2\alpha \end{aligned} \tag{21}$$

(eq 7.6.1 in HN) where α is the significance of the two individual tests.

Multiple testing II

- $\alpha = 0.05$ means that the the Type-1 error probability of the two tests inference procedure is closer to 10%.
- The same potential for *accumulation of inferential errors* arise for sequences of *t-tests*.
- We will return to this issue, and refer §7.6.2 in HN, when we discuss Automatic variable selection later in the course.