Week 5: Estimators and Estimation Methods, Nonlinear Regression, Quantile Regression

Advanced Econometrics 4EK608

Vysoká škola ekonomická v Praze

Outline

Estimators and estimation methods

Notation:

- θ - population parameter
- $(x_1, x_2, \ldots, x_n)$ - random sample of n observation of x
- $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$ is an estimator of θ

Basic notions:

- All estimators posses sampling distribution
  1^{st} moment (mean) $\mathbf{E}(\hat{\theta})$
  2^{nd} moment (variance) $\mathbf{E}[(\hat{\theta} - \mathbf{E}(\hat{\theta}))^2]$
- Estimators $\times$ estimate
- Many estimators exist for a parameter (population mean):

$$\hat{\theta}_1 = \overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\hat{\theta}_2 = \tilde{x} = \frac{1}{2}(x_{max} + x_{min})$$

Estimators and estimation methods

Small sample properties of estimators & definitions:

- **Unbiasedness**: the mean of sampling distribution equals the parameter being estimated
- **Efficiency**: an estimator is efficient if it is unbiased and no other unbiased estimator has a smaller variance. This is usually difficult to prove, that is why we simplify the concept:
    - Relative efficiency
    - Linear unbiased estimators instead of unbiased estimators (linear estimator is linear function of sample observations)

Estimators and estimation methods

Small sample properties of estimators & definitions:

Best Linear Unbiased Estimator (BLUE) is linear, unbiased and no other linear unbiased estimator has a smaller variance. It is not necessarily the best estimator.

- Non-linear estimators can be better
- Biased estimators can have smaller Mean Square Error: sum of variance and the squared bias

Estimators and estimation methods

Large sample properties of estimators & definitions:

- Sampling distribution of an estimator changes with the size of sample.
- Asymptotic distribution for any estimator is that distribution to which the sampling distribution tends as the sample becomes larger. Its 1^{st} and 2^{nd} moments are asymptotic mean and asymptotic variance.
- When the sampling distribution collapses onto a single value when the sample becomes larger, we call this value probability limit. We say estimator converges in probability to that value

Estimators and estimation methods

Large sample properties of estimators & definitions:

- Asymptotic unbiasedness

- **Consistency**
- Unbiased estimators are not necessarily consistent.
- If $\hat{\theta}$ is an unbiased estimator of θ and $var(\hat{\theta} \to 0$ as $n \to \infty$, then $plim(\hat{\theta}) = \theta$.
- Consistent estimators: unibased & their variance shrinks to zero as sample size grows (entire population is used).
  - Minimal requirement for estimator used in statistics or econometrics.
  - If some estimator is not consistent, then it does not help us with estimation of population θ values, even if we have unlimited data.

Estimators and estimation methods

Large sample properties of estimators & definitions:

- Asymptotic efficiency: An estimator is asymptotically efficient if it is asymptotically unbiased and no other asymptotically unbiased estimator has smaller asymptotic variance.
- Asymptotic efficiency is usually difficult to prove, that is why we simplify the concept:
  - Relative asymptotic efficiency
  - Linear asymptotically unbiased estimators instead of asymptotically unbiased estimators

Estimators and estimation methods

Method of moments

- With the method of moments, we simply estimate population moments by corresponding sample moments.
- Under very general conditions, sample moments are consistent estimators of the corresponding population moments, but NOT necessarily unbiased estimators.

Application example 1

Sample covariance is a consistent estimator of population covariance.

Application example 2

OLS estimators we have used for parameters in the CLRM can be derived by the method of moments.

Estimators and estimation methods

Method of moments (MM)

Population moments, stochastic variable X

- $\mathbf{E}(X^r)$: r^{th} population moment about zero
- $\mathbf{E}(X)$: the population mean is the first moment about zero
- $\mathbf{E}[(X-\mathbf{E}(X))^2]$: the population variance is the second moment about the mean

Sample moments, sample observations $(x_1, x_2, \ldots, x_n)$

- $\frac{\sum_{i=1}^{n} x_i^r}{n}$: r^{th} sample moment about zero
- $\frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}$: sample mean is the first moment about zero
- $\frac{\sum_{i=1}^{n} (x_i-\overline{x})^2}{n-1}$: sample variance is the second sample moment about the mean

Estimators and estimation methods

- In a LRM: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$, the $k+1$ parameters are **OLS**-estimated by minimizing:

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik} \right)^2 \tag{1}$$

- 1^{st} order conditions for (1), plus assumptions $E(u) = 0$ and $E(x_j \cdot u) = 0$, can be combined into a **MM** estimator:

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik} \right) = 0$$
$$\sum_{i=1}^{n} x_{i1} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik} \right) = 0$$
$$\cdots$$
$$\sum_{i=1}^{n} x_{ik} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik} \right) = 0$$

Estimators and estimation methods

MM - summary

- MM is robust to differences in “specification” of the data generating process (DGP). $\rightarrow$ i.e. sample mean or sample variance estimate their population counterparts (assuming they exist) regardless of DGP.
- MM is free from distributional assumptions.
- “Cost” of this approach: if we know the specific distribution of a DGP, MM does not make use of such information $\rightarrow$ inefficient estimates.
- Alternative approach: method of maximum likelihood utilizes distributional information and is more efficient (provided this information is valid).

Estimators and Estimation Methods

Maximum likelihood estimator

Single θ parameter case:

- 1^{st} step: deriving a likelihood function $L = L(\theta, x_1, x_2, ..., x_n)$, where x_i is observation, θ is parameter of the distribution.
- 2^{nd} step: finding maximum of L with respect to θ, that maximum is $\tilde{\theta} = \theta_{MLE}$

With more parameters: $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$

$$L = L(\theta_1, \theta_2, ...\theta_m, x_1, x_2, ..., x_n)$$

We find MLEs of the m parameters by partially differentiating the likelihood function L with respect to each θ and then setting all the partial derivatives obtained to zero.

Estimators and estimation methods

- MLE is only possible if we know the form of the probability distribution function for the population.
- MLEs possess the large sample properties of consistency and asymptotic efficiency. There is no guarantee that they possess any desirable small-sample properties.
- Under CLRM assumptions, MLE estimator are identical to OLS estimators.
- **Identification:** The parameter vector $\boldsymbol{\theta}$ is identified (estimable), if for two vectors, $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}$ and for some data observations $\boldsymbol{x}$, $L(\boldsymbol{\theta}^*|\boldsymbol{x}) \neq L(\boldsymbol{\theta}|\boldsymbol{x})$.

Estimators and estimation methods

Maximum likelihood estimation of CLRM parameters:

CLRM: $y_i = \alpha + \beta x_i + \varepsilon_i$ $\qquad$ $\mathbf{E}(y_i) = \alpha + \beta x_i$

$$var(y_i) = var(\varepsilon_i) = \sigma^2$$

Probability density function for normal distribution:

$f(X) = (2\pi\sigma^2)^{-0.5}\, exp[-(x-\mu)^2/2\sigma^2]$ where x is a general random variable

For each y_i

$f(y_i) = (2\pi\sigma^2)^{-0,5}\, exp[-(y_i - \mathbf{E}(y_i))^2/2\sigma^2]$

$$L = f(y_1) \cdot f(y_2) \cdot \ldots \cdot f(y_n)$$

Estimators and estimation methods

Log-likehood function:

$$LL = \sum_{i=1}^{n} log[f(y_i)] =$$
$$= \sum_{i=1}^{n} \left\{ -\frac{1}{2} log(2\pi) - \frac{1}{2} log(\sigma^2) - \frac{1}{2\sigma^2} [y_i - \boldsymbol{E}(y_i)]^2 \right\} =$$
$$= -\frac{n}{2} log(2\pi) - \frac{n}{2} log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} [y_i - \boldsymbol{E}(y_i)]^2$$

max LL is for min $\sum_{i=1}^{n} [y_i - \mathbf{E}(y_i)]^2$
$\Rightarrow$ MLE estimators $\tilde{\alpha}$, $\tilde{\beta}$ are identical to OLS estimators $\hat{\alpha}$, $\hat{\beta}$

Nonlinear regression: linear vs. nonlinear models

Linear model:

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + \varepsilon_i$$
$$y_i = f_1(x_{i1})\beta_1 + f_2(x_{i2})\beta_2 + \cdots + \varepsilon_i$$

Conditional mean function $\mathbf{E}[y|\boldsymbol{x}, \boldsymbol{\beta}] = \boldsymbol{x}'\boldsymbol{\beta}$

Nonlinear models:
Linear model is a special case of the nonlinear model:

$$y_i = h(x_{i1}, x_{i2}, \ldots, x_{ip}; \beta_1, \beta_2, \ldots, \beta_K) + \varepsilon_i$$

Conditional mean function $\mathbf{E}[y|\boldsymbol{x}, \boldsymbol{\beta}] = h(\boldsymbol{x}, \boldsymbol{\beta})$
$\partial\mathbf{E}[y|\boldsymbol{x}], \boldsymbol{\beta}]/\partial\boldsymbol{x}$ is no longer equal to $\boldsymbol{\beta}$,
then how should $\boldsymbol{\beta}$ be interpreted?

For nonlinear models, nonlinear LS have been developed.

Nonlinear regression: linear vs. nonlinear models

Assumptions (comparison with the linear case)

- Functional form
- Identifiability × full rank assumption

The parameter vector in the model is identified (estimable) if there is no nonzero parameter $\boldsymbol{\beta}_0 \neq \boldsymbol{\beta}$ such that $h(\boldsymbol{x}_i, \boldsymbol{\beta}_0) = h(\boldsymbol{x}_i, \boldsymbol{\beta})$ for all $\boldsymbol{x}_i$.

- Zero mean of the disturbances
- Homoscedasticity and nonautocorrelation

Nonlinear regression: linear vs. nonlinear models

- Data Generation Process

The data generating process for $\boldsymbol{x}_i$ is assumed to be a well-behaved population such that first and second moments of the data can be assumed to converge to fixed, finite population counterparts. The crucial assumption is that the process generating $\boldsymbol{x}_i$ is strictly exogenous to that generating ε_i

- Underlying probability model

There is a well-defined probability distribution generating ε_i. At this point, we assume only that this process produces a sample of uncorrelated, identically (marginally) distributed random variables ε_i with mean zero and variance σ^2 conditioned on $h(\boldsymbol{x}_i, \boldsymbol{\beta})$. Thus, at this point, our statement of the model is **semi-parametric**.

Nonlinear Regression: Nonlinear Least Squares

- Minimization of $S(\boldsymbol{\beta}) = \sum[y_i - h(\boldsymbol{x}_i, \boldsymbol{\beta})]^2$

Using standard procedure, we can get k first order conditions.

- In the context of the linear model, the **orthogonality condition** $E[\boldsymbol{x}_i, \varepsilon_i] = 0$ produces least squares as a **GMM estimator** for the linear model. The orthogonality condition is that the regressors and the disturbance in the model are uncorrelated.

Nonlinear regression: nonlinear least squares

- In a similar way, the first order conditions from above are also moment conditions and this defines the nonlinear least squares estimator as a GMM estimator. This - if necessary assumptions (and some other conditions) are fulfilled - allows to deduce that the NLS estimator has good large sample properties: consistency and asymptotic normality.
- Hypothesis testing: The principal testing procedure is the Wald test, which relies on the consistency and asymptotic normality of the estimator - large sample results. The F test relies on normally distributed disturbances, so in the nonlinear case, where we rely on large-sample results, the Wald statistic will be the primary inference tool. **Lagrange multiplier tests** for the general case can also be constructed.

Nonlinear regression: computing NLS estimates

For nonlinear models, a closed-form solution usually does not exist.

- Most of the nonlinear maximization problems are solved by an **iterative algorithm**.
- The most commonly used of iterative algorithms are **gradient methods**.
- The template for most gradient methods in common use is the **Newton's method**.
- Look at your software packages which methods are available for computing NLS estimates.

Nonlinear regression: computing NLS estimates

Examples 7.4 & 7.8 (Greene):

Analysis of a Nonlinear Consumption Function

NLS with starting values equal to 0
NLS with starting values equal to the parameters from the OLS estimation (c(3) equal to 1).

Depednent Variable: REALCONS
Method: Least Squares (Marquard - EViews legacy)
Date: 09/19/16 Time 16:31
Sample 1950Q1 2000Q4
Included observations: 204
REALCONS=C(1)+C(2)*REALDPI

	Coeficient	Std.Error	t-Statistic	Prob.
C(1)	-80.35475	14.30585	-5.616915	0.0000
C(2)	0.921686	0.003872	238.0540	0.0000

R-squared	0.996448	Mean dependent var	2999.436
Adjusted R-squared	0.996431	S.D. dependent var	1459.707
S.E. of regression	87.20983	Akaike info criterion	11.78427
Sum squared resid	1536322	Schwarz criterion	11.81680
Log likelihood	-1199.995	Hannan-Quinn criter.	11.79743
F-statistics	56669.72	Durbin-Watson stat	0.092048
Prob(F-statistics)	0.000000		

Nonlinear regression: computing NLS estimates

Examples 7.4 & 7.8 (Greene): Analysis of a Nonlinear Consumption Function

Depednent Variable: REALCONS
Method: Least Squares (Marquard - EViews legacy)
Sample 1950Q1 2000Q4 Included observations: 204
Convergence achieved after 200 iterations
REALCONS=C(1)+C(2)*REALDPI^C(3)

	Coeficient	Std.Error	t-Statistic	Prob.
C(1)	458.7991	22.50140	20.38980	0.0000
C(2)	0.100852	0.010910	9.243667	0.0000
C(3)	1.244827	0.012055	103.2632	0.0000
R-squared	0.998834	Mean dependent var		2999.436
Adjusted R-squared	0.998822	S.D. dependent var		1459.707
S.E. of regression	50.09460	Akaike info criterion		10.68030
Sum squared resid	504403.2	Schwarz criterion		10.72910
Log likelihood	-1086.391	Hannan-Quinn criter.		10.70004
F-statistics	86081.29	Durbin-Watson stat		0.295995
Prob(F-statistics)	0.000000			

Nonlinear regression: computing NLS estimates

Examples 7.4 & 7.8 (Greene):
Analysis of a Nonlinear Consumption Function

Depednent Variable: REALCONS
Method: Least Squares (Marquard - EViews legacy)
Sample 1950Q1 2000Q4 Included observations: 204
Convergence achieved after 80 iterations
REALCONS=C(1)+C(2)*REALDPI^C(3)

	Coeficient	Std.Error	t-Statistic	Prob.
C(1)	458.7989	22.50149	20.38971	0.0000
C(2)	0.100852	0.010911	9.243447	0.0000
C(3)	1.244827	0.012055	103.2632	0.0000
R-squared	0.998834	Mean dependent var		2999.436
Adjusted R-squared	0.998822	S.D. dependent var		1459.707
S.E. of regression	50.09460	Akaike info criterion		10.68030
Sum squared resid	504403.2	Schwarz criterion		10.72910
Log likelihood	-1086.391	Hannan-Quinn criter.		10.70004
F-statistics	86081.28	Durbin-Watson stat		0.295995
Prob(F-statistics)	0.000000			

Quantile regression

- Quantile regression provides estimates of the relationship between regressors and a specified quantile of the dependent variable.
- The (linear) quantile model can be defined as $Q[y|\boldsymbol{x}, q] = \boldsymbol{x}\boldsymbol{\beta}$, such that $\text{Prob}[y \leq \boldsymbol{x}\boldsymbol{\beta}_q|\boldsymbol{x}] = q$, $0 < q < 1$.
- One important special case of quantile regression is the least absolute deviations (LAD) estimator, which corresponds to fitting the conditional median of the response variable ($q = \frac{1}{2}$).
- LAD estimator can be also motivated as a robust (to outliers) alternative to LS.

Quantile regression

- The LAD estimator is the solution to the optimization problem: $\underset{\hat{\beta}}{min} \sum |y_i - x_i\hat{\beta}|$
- The LAD estimator's history predates least squares (which itself was proposed over 200 years ago). It has seen little use in econometrics, primarily for the same reason that Gauss's method (LS) supplanted LAD at its origination; LS is vastly easier to compute.
- Look at your software packages which methods are available for quantile regression.
- It can be of some interest that the original approaches used linear programming for finding the estimate (Koenkerr and Bassett (around 1980)).

Quantile regression

Examples 7.9 (Greene): Cobb-Douglass Production Function
OLS → Standardized residuals indicate two outliers → LAD

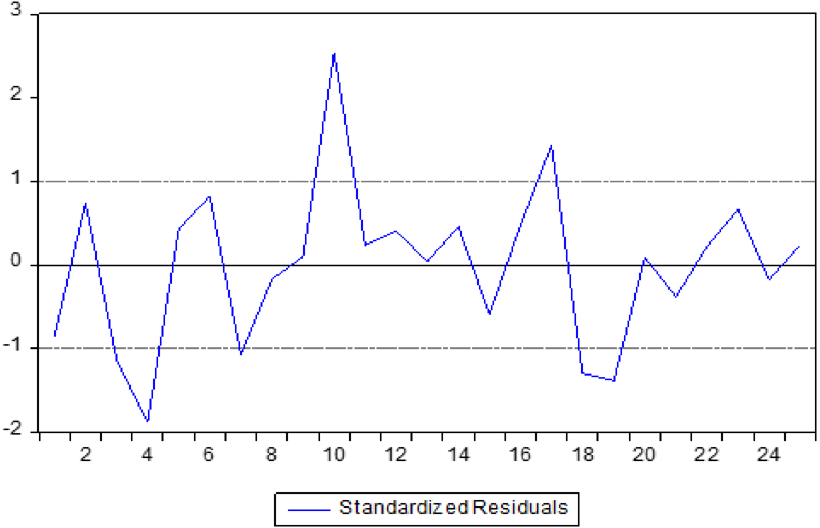
Depednent Variable: LNYN
Method: Least Squares
Sample 1 25
Included observations: 25

Variable	Coeficient	Std.Error	t-Statistic	Prob.
C	2.293263	0.107183	21.39582	0.0000
LNKN	0.278982	0.080686	3.457639	0.0022
LNLN	0.927312	0.012055	9.431359	0.0000

R-squared	0.959742	Mean dependent var	0.771734
Adjusted R-squared	0.956082	S.D. dependent var	0.899306
S.E. of regression	0.188463	Akaike info criterion	-0.387663
Sum squared resid	0.781403	Schwarz criterion	-0.241398
Log likelihood	7.845786	Hannan-Quinn criter.	-0.347095
F-statistics	262.2396	Durbin-Watson stat	1.937830
Prob(F-statistics)	0.000000		

Quantile regression

Examples 7.9 (Greene): Cobb-Douglass Production Function



Quantile Regression

Examples 7.9 (Greene): Cobb-Douglass Production Function (results differ from the textbook results)

Depednent Variable: LNYN Method: Quantile Regression (Median)
Sample 1 25 Included observations: 25
Huber Sandwich Standard Errors & Covariance
Sparsity method: Kemel (Epanechnikov) using residuals
Bandwidth method: Hall-Sheather, bw=0.33227
Estimation successfully identifies unique optimal solution

Variable	Coeficient	Std.Error	t-Statistic	Prob.
C	2.275038	0.179268	12.69071	0.0000
LNKN	0.260365	0.122447	2.126351	0.0449
LNLN	0.927243	0.152593	6.076572	0.0000

Pseudo R-squared	0.794575	Mean dependent var	0.771734
Adjusted R-squared	0.775900	S.D. dependent var	0.899306
S.E. of regression	0.190505	Objective	1.627051
Quantile dependent va...	0.966677	Restr. objective	7.920415
Sparsity	0.594465	Quasi-LR statistic	84.69274
Prob(Quasi-LR stat)	0.000000		

Quantile regression

Examples 7.10 (Greene): Income Elasticity of Credit Cards Expenditure

OLS → LAD → Income Elasticity for 10 Deciles

Depednent Variable: LOGSPEND
Method: Least Squares
Date: 09/15/16 Time 13:53
Sample (adjusted): 3 13443
Included observations: 10499 after adjustments

Variable	Coeficient	Std.Error	t-Statistic	Prob.
C	-3.055807	0.239699	-12.74852	0.0000
LOGINC	1.083438	0.032118	33.73296	0.0000
AGE	-0.017364	0.001348	-12.88069	0.0000
ADEPCNT	-0.044610	0.010921	-4.084857	0.0000
R-squared	0.100572	Mean dependent var		4.728778
Adjusted R-squared	0.100315	S.D. dependent var		1.404820
S.E. of regression	1.332496	Akaike info criterion		3.412366
Sum squared resid	18634.35	Schwarz criterion		3.415131
Log likelihood	-17909.21	Hannah-Quinn criter.		3.413300
F-statistic	391.1750	Durbin-Watson stat		1.888912
Prob(F-statistic)	0.000000			

Quantile regression

Examples 7.10 (Greene): Income Elasticity of Credit Cards Expenditure

Depednent Variable: LOGSPEND Method: Quantile Regression (Median)
Sample (adjusted): 3 13443 Included observations: 10499 after adjustments
Huber Sandwich Standard Errors & Covariance
Sparsity method: Kemel (Epanechnikov) using residuals
Bandwidth method: Hall-Sheather, bw=0.04437
Estimation successfully identifies unique optimal solution

Variable	Coeficient	Std.Error	t-Statistic	Prob.
C	-2.803756	0.233534	-12.00577	0.0000
LOGINC	1.074928	0.030923	34.76139	0.0000
AGE	-0.016988	0.001530	-11.10597	0.0000
ADEPCNT	-0.049955	0.011055	-4.518599	0.0000
Pseudo R-squared	0.058243	Mean dependent var		4.728778
Adjusted R-squared	0.057974	S.D. dependent var		1.404820
S.E. of regression	1.346476	Objective		5096.818
Quantile dependent va...	4.941583	Restr. objective		5412.032
Sparsity	2.659971	Quasi-LR statistic		948.0224
Prob(Quasi-LR stat)	0.000000			

Quantile regression

Examples 7.10 (Greene): Income Elasticity of Credit Cards Expenditure