Block 1
Repetition from BSc courses
LRM estimators & non-linear extensions
Predictions from regression models

Advanced Econometrics 4EK608

Vysoká škola ekonomická v Praze

Outline

1. Estimation methods, predictions from a model
   - Ordinary least squares
   - General properties of estimators
   - Method of moments
   - Maximum likelihood estimator
2. Non-linear extensions to LRM, quantile regression
   - Non-linear regression models
   - Quantile regression
3. Predictions from a regression model
   - Predictions from a CLRM (repetition from BSc courses)
   - Predictions: general features, kFCV, Variance vs. Bias

Linear regression model (LRM) and OLS estimation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

LRM assumptions (for OLS estimation):
(Notation follows Greene, Econometric analysis, 7^{th} ed.)

- A1 **Linearity:** $y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i$
  LRM describes linear relationship between y_i and $\boldsymbol{x}_i$.
- A2 **Full rank:** Matrix $\boldsymbol{X}$ is an $n \times K$ matrix with rank K.
  Columns of $\boldsymbol{X}$ are linearly independent and $n \geq K$.
- A3 **Exogeneity of regressors:** $E[\varepsilon_i|\boldsymbol{X}] = 0$ (strict form).
  If relaxed to contemporaneous form in TS: $E[\varepsilon_t|\boldsymbol{x}_t] = 0$.
  Law of iterated expectations: $E[\varepsilon_i|\boldsymbol{X}] = 0 \;\Rightarrow\; E[\varepsilon] = 0$.
  Also, we assume disturbances convey no information on each other: $E[\varepsilon_i|\varepsilon_1, \ldots, \varepsilon_{i-1}, \varepsilon_{i+1}, \ldots, \varepsilon_n] = 0$.

Linear regression model (LRM) and OLS estimation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

LRM assumptions (continued):

A4 **Homoscedastic & nonautocorrelated disturbances:**

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \sigma^2 \boldsymbol{I}$$

Homoscedasticity: $\text{var}[\varepsilon_i | \boldsymbol{X}] = \sigma^2, \qquad \forall\ i = 1, \ldots, n.$
Independent disturbances: $\text{cov}[\varepsilon_t, \varepsilon_s | \boldsymbol{X}] = 0, \qquad \forall\ t \neq s.$
GARCH models (i.e. $\text{var}[\varepsilon_t | \varepsilon_{t-1}] = \sigma^2 + \alpha\varepsilon_{t-1}$) do not violate the conditional variance assumption, but $\text{var}[\varepsilon_t | \varepsilon_{t-1}] \neq \text{var}[\varepsilon_t]$.

A5 **DGP of $\boldsymbol{X}$:** Variables in $\boldsymbol{X}$ may be fixed or random.

A6 **Normal distribution of disturbances:**

$$\varepsilon | \boldsymbol{X} \sim N[\boldsymbol{0}, \sigma^2 \boldsymbol{I}].$$

Ordinary least squares (OLS)

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

The least squares estimator is unbiased (given A1 – A3):

$$\hat{\boldsymbol{\beta}} = \boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \;=\; \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon},$$

$$\text{take expectations, iterating over } \boldsymbol{X}:$$

$$E[\boldsymbol{b}|\boldsymbol{X}] = \boldsymbol{\beta} + E[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}|\boldsymbol{X}] \;=\; \boldsymbol{\beta}.$$

Variance of the least squares estimator (A1 – A4):

$$\begin{aligned}
\operatorname{var}[\boldsymbol{b}|\boldsymbol{X}] &= E[(\boldsymbol{b}-\boldsymbol{\beta})(\boldsymbol{b}-\boldsymbol{\beta})'|\boldsymbol{X}] \\
&= E[\boldsymbol{A}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\boldsymbol{A}'|\boldsymbol{X}] \quad \text{where} \quad \boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.
\end{aligned}$$

Normal distribution of the least squares estimator (A1 – A6):

$$\boldsymbol{b}|\boldsymbol{X} \sim N[\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}].$$

Estimators and estimation methods

- LRM is not the only type of regression model.
- OLS is not the only useful estimator.
- Let's approach estimators and their properties more generally.

  (again, notation follows Greene, Econometric analysis.)

Estimators and estimation methods

Notation/definitions:

- $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{nj})'$ - random sample of n observations.
- $\boldsymbol{\theta}$ - population parameter [unknown parameter(s)]
- $f(\boldsymbol{x}_j, \boldsymbol{\theta})$: probability distribution function
- $\hat{\boldsymbol{\theta}}$ is some estimator of $\boldsymbol{\theta}$

Basic notions:

- All estimators have sampling distributions
  mean: $E(\hat{\theta})$
  variance: $E[(\hat{\theta} - E(\hat{\theta}))^2]$, etc.
- Estimators $\times$ estimate
- Generally, many estimators exist for a given parameter.
  Population mean example:

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

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

Estimators and estimation methods

Properties of estimators - classification:

- **Unbiasedness**: can be described as $E(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$.
  Rarely useful in finite (small) sample context. Even asymptotically (large sample), discussion would be directed towards consistency (far more desirable feature).

- **Consistency:** $\text{plim}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$.
  Holds if $\hat{\boldsymbol{\theta}}$ is an unbiased estimator of $\boldsymbol{\theta}$ and $\text{plim}(\text{var}(\hat{\boldsymbol{\theta}})) = \mathbf{0}$ i.e. $[\text{var}(\hat{\boldsymbol{\theta}}) \to \mathbf{0}$ as $n \to \infty]$.
  - Consistent estimators: unbiased & 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 provide estimates of population θ values, even with unlimited data.
  - Unbiased estimators are not necessarily consistent.

Estimators and estimation methods

Properties of estimators - classification:

- **Efficiency, asymptotic efficiency:** an estimator is efficient if it is unbiased and no other unbiased estimator has a smaller variance. Often difficult to prove, we usually simplify the concept to *relative efficiency* (e.g.: efficiency with respect to linear unbiased estimators, etc.). Asymptotic efficiency: Holds for an estimator that is asymptotically unbiased and no other asymptotically unbiased estimator has smaller asymptotic variance.
- **Normality, asymptotic normality:** basis for most statistical inference performed with common estimators.

Estimators and estimation methods

Extremum estimator: obtained as the optimizer of some criterion function $q(\boldsymbol{\theta}|\mathbf{data})$. Most common estimators:

$$
\begin{aligned}
\text{LS} \quad \hat{\boldsymbol{\theta}}_{LS} &= \operatorname{argmax} \left[-\tfrac{1}{n} \sum_{i=1}^{n} (y_i - h(\boldsymbol{x}_i, \boldsymbol{\theta}_{LS}))^2 \right], \\
\text{ML} \quad \hat{\boldsymbol{\theta}}_{ML} &= \operatorname{argmax} \left[\tfrac{1}{n} \sum_{i=1}^{n} \log f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}_{ML}) \right], \\
\text{GMM} \quad \hat{\boldsymbol{\theta}}_{GMM} &= \operatorname{argmax} \left[-\overline{\boldsymbol{m}}(\mathbf{data}, \boldsymbol{\theta}_{GMM})' \boldsymbol{W} \, \overline{\boldsymbol{m}}(\mathbf{data}, \boldsymbol{\theta}_{GMM}) \right],
\end{aligned}
$$

where $h(\cdot)$ is a function (linear/non-linear $\rightarrow$ OLS/NLS), $\overline{\boldsymbol{m}}$ denotes sample moments and $\boldsymbol{W}$ is a convenient positive definite matrix.

LS and ML estimators belong to a class of **M estimators** ("M" for maximum likelihood-type).

Estimators and estimation methods

Assumptions for asymptotic properties of extremum estimators:

1. **Parameter space:** must be convex and the parameter vector that is the object of estimation must be point in its interior. Gaps and nonconvexities in parameter spaces would generally collide with estimation algorithms (settings such as $\sigma^2 > 0$ are OK).
2. **Criterion function:** must be concave in the parameters (concave in the neighborhood of the true parameter vector). Criterion functions need not be globally concave. In such situation, there may be multiple optima (often associated with poor model specification).

Estimators and estimation methods

Assumptions for asymptotic properties of extremum estimators:

3. **Identifiability of parameters:** has a relatively complex technical definition (anything like "true parameters $\boldsymbol{\theta}_0$ are identified if..." is problematic - leads to a paradox if condition is not met). Simple way to secure identification:
    - **LS:** for a given set of $\boldsymbol{x}_i$ observations, any two different parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ must lead to different conditional mean function ($\hat{y}_i$).
    - **ML:** For any data vector $(y_i, \boldsymbol{x}_i)$ and two parameter vectors $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, it must be possible to produce different values of the density function $f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})$. Note: identifiability does not rule out the $f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) = f(y_i|\boldsymbol{x}_\ell, \boldsymbol{\theta}), \quad i \neq \ell$ situation (A2 full rank condition must be observed).
    - **GMM:** sufficient condition for identification: $E[\overline{\boldsymbol{m}}(\mathbf{data}, \boldsymbol{\theta})] \neq \mathbf{0}$ if $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$.

Assumptions for asymptotic properties of extremum estimators:

4. **Behavior of the data:** Grenander conditions for well-behaved data:
   - G1 For each $\boldsymbol{x}_k$ column of $\boldsymbol{X}$ and $d_{nk}^2 = \boldsymbol{x}_k'\boldsymbol{x}_k$, it must hold that: $\lim_{n\to\infty} d_{nk}^2 = +\infty$. Sum of squares continue to grow with sample size ($\boldsymbol{x}_k$ does not degenerate into a series of 0).
   - G2 The $\lim_{n\to\infty} x_{ik}^2/d_{nk}^2 = 0$ for all $i = 1, 2, \ldots, n$. Single observations become less important as sample size grows. No single observation will dominate $\boldsymbol{x}_k'\boldsymbol{x}_k$.
   - G3 Let $\boldsymbol{C}_n$ be sample correlation matrix of the columns in $\boldsymbol{X}$ (excluding the intercept, if present). Then $\lim_{n\to\infty} \boldsymbol{C}_n = \boldsymbol{C}$ where $\boldsymbol{C}$ is positive definite. This implies that the full rank condition for $\boldsymbol{X}$ (A2) is not asymptotically violated.

Theorem: Consistency of M estimators

If:

(a) the parameter space is convex and the true parameter vector is a point in its interior,

(b) the criterion function is concave,

(c) the parameters are identified by the criterion function,

(d) the data are well behaved,

then the M estimator converges in probability to the true parameter vector.

Estimators and estimation methods

Theorem: Asymptotic normality of M estimators

If:

(a) $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0$ is a point in the interior of the parameter space $\boldsymbol{\Theta}$,

(b) $q(\boldsymbol{\theta}|\mathbf{data})$ is concave and twice continuously differentiable in $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$,

(c) $\sqrt{n}\,[\partial q(\boldsymbol{\theta}_0|\mathbf{data})/\partial \boldsymbol{\theta}_0] \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Phi})$,

(d) $\lim\limits_{n\to\infty} \Pr\left[|(\partial^2 q(\boldsymbol{\theta}|\mathbf{data})/\partial\theta_k\partial\theta_m) - h_{km}(\boldsymbol{\theta})| > \varepsilon\right] = 0 \; \forall \varepsilon > 0$ for any $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}$; $h_{km}(\boldsymbol{\theta})$ is a continuous finite valued function of $\boldsymbol{\theta}$,

(e) the matrix of elements $\boldsymbol{H}(\boldsymbol{\theta})$ is nonsingular at $\boldsymbol{\theta}_0$,

then $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N\{\mathbf{0}, [\boldsymbol{H}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{\Phi}\boldsymbol{H}^{-1}(\boldsymbol{\theta}_0)]\}$.

where $\boldsymbol{\Phi}$ is a variance-covariance matrix,
and $\boldsymbol{H}(\boldsymbol{\theta}_0) = \partial^2 q(\boldsymbol{\theta}|\mathbf{data})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ is a Hessian (evaluated at $\boldsymbol{\theta}_0$).

Method of moments

- Method of moments (MM)
- Generalized method of moments (GMM)

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.

Method of moments (MM)

Population moments for a stochastic variable X

- $E(X^r)$: r^{th} population moment about zero
- $E(X)$: the population mean: 1^{st} sample moment about zero
- $E[(X - E(X))^2]$: the population variance is the second moment about the mean

Sample moments for 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

Method of moments

- For MM, the usual linear model assumption (concerning 1^{st} population moment) $E[\boldsymbol{x}_i \varepsilon_i] = \mathbf{0}$ implies orthogonality condition:

$$\text{cov}[\boldsymbol{x}_i, \varepsilon_i] = \mathbf{0} \quad \text{i.e.} \quad E[\boldsymbol{x}_i(y_i - \boldsymbol{x}_i' \boldsymbol{\beta}] = \mathbf{0},$$

may be transformed in **population moment equation:**

$$E\left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})\right] = E\left[\overline{\boldsymbol{m}}(\boldsymbol{\beta})\right] = \mathbf{0}\,,$$

and corresponding sample (empirical) moment equation:

$$\left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}})\right] = \overline{\boldsymbol{m}}(\hat{\boldsymbol{\beta}}) = \mathbf{0}.$$

Method of moments

The equation form of MM empirical equations can be produced as:

$$\frac{1}{n}\sum_{i=1}^{n} \left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_K x_{iK}\right) = 0$$

$$\frac{1}{n}\sum_{i=1}^{n} x_{i1}\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_K x_{iK}\right) = 0$$

$$\cdots$$

$$\frac{1}{n}\sum_{i=1}^{n} x_{iK}\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_K x_{iK}\right) = 0$$

- The $\frac{1}{n}$ element can be removed here.
- This is a system of K equations with K unknown parameters.
- Equivalent to 1^{st} order conditions for the OLS estimator:

$$\min_{\hat{\boldsymbol{\beta}}} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_K x_{iK}\right)^2$$

Generalized method of moments

- GMM is a very general class of estimators, includes most of other estimators as a special case (IVR, simultaneous equations, Arellano-Bond estimator for dynamic panels).
- For single equation linear models, GMM may be conveniently described using the instrumental variable case:

  For the LRM $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i$,

  - we abandon the assumption $E[\boldsymbol{x}_i'\varepsilon_i] = 0$ and
  - we replace it by $E[\boldsymbol{z}_i'\varepsilon_i] = 0$.
  - Hence, columns of $\boldsymbol{X}$ $(n \times K)$ are potentially endogenous and $\boldsymbol{Z}$ $(n \times L)$ is a matrix of exogenous instruments.

Generalized method of moments

- GMM equation can be cast by analogy to the MM case: we start by $E[\boldsymbol{z}_i \varepsilon_i] = \mathbf{0}$, which implies orthogonality condition:

$$\text{cov}[\boldsymbol{z}_i, \varepsilon_i] = \mathbf{0} \quad \text{i.e.} \quad E[\boldsymbol{z}_i(y_i - \boldsymbol{x}_i' \boldsymbol{\beta}] = \mathbf{0}.$$

  and **population moment equation:**

$$E\left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})\right] = E\left[\overline{\mathbf{m}}(\boldsymbol{\beta})\right] = \mathbf{0}\,,$$

  and corresponding sample (empirical) moment equation:

$$\left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}})\right] = \overline{\mathbf{m}}(\hat{\boldsymbol{\beta}}) = \mathbf{0}.$$

Generalized method of moments

The equation form of GMM empirical equations can be produced as:

$$\frac{1}{n}\sum_{i=1}^{n} \left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_K x_{iK}\right) = 0$$

$$\frac{1}{n}\sum_{i=1}^{n} z_{i1}\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_K x_{iK}\right) = 0$$

$$\cdots$$

$$\frac{1}{n}\sum_{i=1}^{n} z_{iL}\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_K x_{iK}\right) = 0$$

- First column of $\boldsymbol{Z}$ is assumed to be a vector of ones (same as for $\boldsymbol{X}$).
- For $\boldsymbol{Z} = \boldsymbol{X}$ as a special case, the above equations are identical to MM (shown previously) and the solution is identical to the OLS estimator: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$.
- For $\boldsymbol{Z} \neq \boldsymbol{X}$, where $\boldsymbol{Z}$ is $(n \times L)$ and $\boldsymbol{X}$ is $(n \times K)$, three identification possibilities have to be considered.

Identification of GMM equations

1 **Underidentified:** with $L < K$, there are fewer moment equations than unknown parameters (β_j). Without additional information (parameter restrictions), there is no solution to the system of GMM equations.

2 **Exactly identified:** for $L = K$, single solution exists:

$$\left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}})\right] = \overline{\mathbf{m}}(\hat{\boldsymbol{\beta}}) = \mathbf{0},$$

can be conveniently re-written as:

$$\overline{\mathbf{m}}(\hat{\boldsymbol{\beta}}) = \left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{y}\right) - \left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\hat{\boldsymbol{\beta}} = \mathbf{0}$$

and the solution yields the familiar IV estimator:

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

Identification of GMM equations (continued)

3. With $L > K$, there is no unique solution to the equation system $\overline{\mathbf{m}}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$.

   One intuitite solution is the "least squares approach":

$$\min_{\boldsymbol{\beta}} \left(\overline{\mathbf{m}}(\hat{\boldsymbol{\beta}})' \overline{\mathbf{m}}(\hat{\boldsymbol{\beta}}) \right)$$

   Through the first order conditions, we obtain a GMM estimator as

$$\hat{\boldsymbol{\beta}} = \left[(\boldsymbol{X}'\boldsymbol{Z})(\boldsymbol{Z}'\boldsymbol{X}) \right]^{-1} (\boldsymbol{X}'\boldsymbol{Z})\boldsymbol{Z}'\boldsymbol{y}.$$

Generalized method of moments

GMM - consistency conditions

- **Convergence of the moments:** Empirical (sample) moments converge in probability to their population counterparts. DGP meets the conditions for LLN. $\overline{\mathbf{m}}(\boldsymbol{\beta}) = \frac{1}{n}\left(\boldsymbol{Z}'\boldsymbol{y} - \boldsymbol{Z}'\boldsymbol{X}\boldsymbol{\beta}\right) \xrightarrow{p} \mathbf{0}$.
- **Identification:** For any $n \geq K$ and $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$ it holds that $\overline{\mathbf{m}}(\boldsymbol{\beta}_1) \neq \overline{\mathbf{m}}(\boldsymbol{\beta}_2)$. Three implications:
  - **Order condition:** $L \geq K$. Number of moment equations at least as large as number of parameters.
  - **Rank condition:** matrix $\boldsymbol{G}(\boldsymbol{\beta}) = \partial\,\overline{\mathbf{m}}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}'$ (i.e. $\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}$) is a $L \times K$ matrix with row rank equal to K. Moment conditions are not redundant (implies order condition).
  - **Uniqueness:** unique solution/optimizer exists.
- **Limiting Normal distribution for the sample moments:** Population moments obey central limit theorem (CLT) or some similar variant.

GMM - final remarks & summary

- GMM-based asymptotic covariance matrix of $\hat{\beta}$ is discussed in Greene (Econometric analysis, ch. 13.6) for the classical, heteroscedastic and generalized case (includes TS-based estimation).
- GMM 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.
- GMM is free from distributional assumptions. "Cost" of this approach: if we know the specific distribution of a DGP, GMM 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).

- Maximum likelihood estimator (MLE)
- Normal distribution & MLE

Maximum likelihood estimator

Maximum likelihood estimator – single parameter

For a stochastic variable y with a known distribution, described by a single θ parameter:

- $f(y|\theta)$ is the pdf of y, conditioned on parameter θ.
- For n *iid* observations, joint density of this process:

  $$f(y_1, y_2, \ldots, y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta) = L(\theta|\boldsymbol{y})$$

  is the likelihood function.
- We estimate θ by maximizing $L(\theta|\boldsymbol{y})$ with respect to the parameter (1^{st} order conditions). Solution (MLE) often denoted as $\hat{\theta}_{\text{ML}}$.
- For maximization (MLE), it is usually simpler to work with a log-transformed likelihood function:

  $$\log L(\theta|\boldsymbol{y}) = \sum_{i=1}^{n} \log f(y_i|\theta).$$

Maximum likelihood estimator

MLE – Poisson distribution example

- Consider 10 *iid* observations from a Poisson distribution: $\boldsymbol{y}' = (5, 0, 1, 1, 0, 3, 2, 3, 4, 1)$.
- The pdf: $f(y_i|\lambda) = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$.
- Hence $(n = 10)$: $L(\lambda|\boldsymbol{y}) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} = \frac{e^{-10\lambda}\lambda^{\sum_{i=1}^{10} y_i}}{\prod_{i=1}^{10} y_i!}$.
- *LL* (general): $\log L(\lambda|\boldsymbol{y}) = -n\lambda + \log\lambda \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \log(y_i!)$,
- 1^{st} order condition: $\frac{\partial \log L(\lambda|\boldsymbol{y})}{\partial \lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} y_i = 0$.
- From 1^{st} order condition: $\hat{\lambda}_{\text{ML}} = \overline{y}_n$.
  For our empirical example, $\hat{\lambda}_{\text{ML}} = 2$.

Maximum likelihood estimator

Maximum likelihood estimator – vector of parameters

- $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$
- $L = L(\theta_1, \theta_2, ...\theta_m | y_1, y_2, ..., y_n)$
- We find MLEs of the m parameters by partially differentiating the likelihood function L (often, log L is used) with respect to each θ and then setting all the partial derivatives obtained to zero.

Maximum likelihood estimator

MLE – Normal distribution

- $L(\boldsymbol{\theta}|\textbf{data}) = L(\boldsymbol{\beta}, \sigma^2|y_i, \boldsymbol{x}_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-\frac{(y_i - \boldsymbol{x}_i \boldsymbol{\beta})^2}{2\sigma^2}}$
- In matrix form, the log likelihood function is:

$$LL(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{X}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

Recall that:

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$$

and

$$\frac{\partial(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{\partial\boldsymbol{\beta}'} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}.$$

Maximum likelihood estimator

MLE – Normal distribution (continued)

$$LL(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}, \boldsymbol{X}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

- 1^{st} order conditions:
  - $\frac{\partial LL}{\partial \boldsymbol{\beta}'} = \frac{1}{2\sigma^2}[2\boldsymbol{X}'\boldsymbol{y} - 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}] = \boldsymbol{0}$

    is solved by:

    $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$
  - $\frac{\partial LL}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})] = 0$

    is solved by:

    $\hat{\sigma}^2 = \frac{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})}{n} = \frac{\boldsymbol{u}'\boldsymbol{u}}{n} = \frac{\text{SSR}}{n}$.

    Note: the MLE estimate $\hat{\sigma}^2$ is biased downwards in small samples, as the unbiased estimate is equal to $\text{SSR}/(n-K)$.

Basic MLE assumptions

- **Parameter space:** Gaps and nonconvexities in parameter spaces would generally collide with estimation algorithms (settings such as $\sigma^2 > 0$ are OK).
- **Identifiability:** The parameter vector $\boldsymbol{\theta}$ is identified (estimable), if for two vectors, $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}$ and for some data observations $\boldsymbol{x}$, $L(\boldsymbol{\theta}^*|\boldsymbol{x}) \neq L(\boldsymbol{\theta}|\boldsymbol{x})$.
- **Well-behaved data:** Laws of large numbers (LLN) apply. Some form of CLT can be applied to the gradient (i.e. for the estimation method).
- **Regularity conditions:** "well behaved" derivatives of $f(y_i|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ (see Greene, chapter 14.4.1).

Maximum likelihood estimator

MLE properties

- **Consistency:** $\text{plim}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}_0 \quad (\boldsymbol{\theta}_0$ is the true parameter)
- **Asymptotic normality** of $\hat{\boldsymbol{\theta}}$
- **Asymptotic efficiency:** $\hat{\boldsymbol{\theta}}$ is asymptotically efficient and achieves the Cramér-Rao lower bound for consistent estimators (see Greene, chapter 14.4.5)
- **Invariance:** MLE of $\boldsymbol{\gamma}_0 = \boldsymbol{c}(\boldsymbol{\theta}_0)$ is $\boldsymbol{c}(\hat{\boldsymbol{\theta}})$ if $\boldsymbol{c}(\boldsymbol{\theta}_0)$ is a continuous and countinuously differentiable function.

  (empirical advantages: we can use reparameterization in MLE, e.g. $\gamma_j = 1/\theta_j \quad \text{or} \quad \theta^2 = 1/\sigma^2$).

MLE - properties of the estimator (Normal distribution):

Under the above assumption, variance-covariance matrix of $\boldsymbol{\theta}$ is the inverse of the Information matrix:

$$\mathrm{var}(\hat{\boldsymbol{\theta}}) = \boldsymbol{I}[\hat{\boldsymbol{\theta}}]^{-1} = \begin{bmatrix} \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{2\sigma^4}{n} \end{bmatrix},$$

where $\boldsymbol{I}[\boldsymbol{\theta}] = -E[\boldsymbol{H}(\boldsymbol{\theta})]$. MLE gives the familiar formula for the variance-covariance matrix of the $\hat{\boldsymbol{\beta}}$: $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, and a simple expression for the variance of $\hat{\sigma}^2$.

- The square root of the diagonal elements of $\boldsymbol{I}[\hat{\boldsymbol{\theta}}]^{-1}$ gives estimates of the standard errors of the parameter estimates.
- We can construct simple z-scores to test the null hypothesis concerning any individual parameter, just as in OLS, but using the normal instead of the t-distribution.

MLE - inference, three classic tests:

Consider MLE of parameter $\boldsymbol{\theta}$ and a test of the hypothesis $H_0 : \boldsymbol{h}(\boldsymbol{\theta}) = \mathbf{0}$. Recall that ML parameter estimates are asymptotically normally distributed.

1. **Likelihood ratio test:** If the restriction $\boldsymbol{h}(\boldsymbol{\theta}) = \mathbf{0}$ is valid, then imposing it should not lead to a large reduction in the log-likelihood function.

$$LR \; = \; -2(LL_U - LL_R) \underset{H_0}{\sim} \chi^2(r),$$

where LL_U is the LL of unconstrained model, LL_R denotes restricted model and r is the number of restrictions imposed. To do this test you have to run two models (one nested) and get the results of both.

Maximum likelihood estimator

MLE - inference, three classic tests:

We have a ML estimate $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)'$,
and test of the hypothesis $H_0 : \boldsymbol{h}(\boldsymbol{\theta}) = \mathbf{0}$,
where $\boldsymbol{h}(\boldsymbol{\theta})$ is a $(r \times 1)$ vector function of $\boldsymbol{\theta}$ (linear/non-linear restrictions, continuous partial derivatives assumed).

2 **Wald test:** If the restriction $\boldsymbol{h}(\boldsymbol{\theta}) = \mathbf{0}$ is valid, then $\boldsymbol{h}(\hat{\boldsymbol{\theta}})$ should be close to zero since MLE is consistent.

$$W = \boldsymbol{h}(\hat{\boldsymbol{\theta}})' \left[\boldsymbol{H}(\hat{\boldsymbol{\theta}}) \, \widehat{\text{cov}}(\hat{\boldsymbol{\theta}}) \, \boldsymbol{H}'(\hat{\boldsymbol{\theta}}) \right]^{-1} \boldsymbol{h}(\hat{\boldsymbol{\theta}}) \quad \underset{H_0}{\sim} \chi^2(r),$$

where $\boldsymbol{H}(\hat{\boldsymbol{\theta}}) = \partial \boldsymbol{h}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}'$ at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ and $\text{rank}(\boldsymbol{H}) = r$.

Maximum likelihood estimator

MLE - inference, three classic tests:

We have a ML estimate $\hat{\boldsymbol{\theta}}_R$ – i.e. ML estimation of the restricted model, under $H_0 : \boldsymbol{h}(\boldsymbol{\theta}) = \mathbf{0}$,

3 **Lagrange multiplier test:** If the restriction is valid, then the restricted estimator should be near the point that maximizes the log-likelihood. Therefore, the slope of the log-likelihood function should be near zero at the restricted estimator. The test is based on the slope of the log-likelihood at the point where the function is maximized subject to the restriction.

$$LM = \left(\frac{\partial \log L(\hat{\boldsymbol{\theta}}_R)}{\partial \hat{\boldsymbol{\theta}}_R} \right)' \boldsymbol{I}[\hat{\boldsymbol{\theta}}_R]^{-1} \left(\frac{\partial \log L(\hat{\boldsymbol{\theta}}_R)}{\partial \hat{\boldsymbol{\theta}}_R} \right) \underset{H_0}{\sim} \chi^2(r),$$

where $-\boldsymbol{I}[\hat{\boldsymbol{\theta}}_R] = \partial^2 LL(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}' \partial \boldsymbol{\theta}$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_R$.

Maximum likelihood estimator

MLE - inference, three classic tests:

- The χ^2 distributions of the three test statistics are asymptotically valid.
- The three tests are asymptotically equivalent, but may differ in small samples:
- $W \;\geq\; LR \;\geq\; LM$.
- Hence, in finite samples, LR rejects H_0 less often than W but more often than LM.
- The test are discussed in ML context, i.e. with known distribution of the variable (ML parameter estimates are asymptotically normally distributed).

Maximum likelihood estimator

MLE – summary

- MLE is only possible if we know the form of the probability distribution function for the population (Normal, Poisson, Negative Binomial, etc.).
- MLE has the large sample properties of consistency and asymptotic efficiency. There is no guarantee of desirable small-sample properties.
- Under CLRM assumptions (A1 – A6), ML estimator is identical to OLS estimator (for $\hat{\boldsymbol{\beta}}$).

Non-linear extensions to LRM, quantile regression

- Non-linear regression models
- Quantile regression

Non-linear regression models

Nonlinear regression model:

$$y_i = h(\boldsymbol{x}_i\,,\,\boldsymbol{\beta}) + \varepsilon_i$$

- Linear model is a special case of the nonlinear model.
  - $y_i \;=\; h(\boldsymbol{x}_i\,,\,\boldsymbol{\beta}) + \varepsilon_i \;=\; \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i$.
  - Linear models: linear in parameters. Definition includes non-linear regressors such as x_i^2, etc.
  - Many nonlinear model may be transformed into linear models (log-transformation)
- For nonlinear models that cannot be transformed into LRM, nonlinear LS (NLS) are available.
- $\partial h(\boldsymbol{x}_i\,,\,\boldsymbol{\beta})/\partial \boldsymbol{x}$ is no longer equal to $\boldsymbol{\beta}$ (interpretation based on estimated model ...)

Nonlinear regression

Assumptions relevant to the nonlinear regression model

1. **Functional form:** The conditional mean function for y_i, given $\boldsymbol{x}_i$ is:

$$\mathbf{E}[y_i|\boldsymbol{x}_i] = h(\boldsymbol{x}_i\,,\,\boldsymbol{\beta})\ , \quad i = 1, 2, \ldots, n$$

2. **Identifiability of model parameters:** 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 $\boldsymbol{x}_i$.

3. **Zero mean of the disturbance:** For $y_i = h(\boldsymbol{x}_i\,,\,\boldsymbol{\beta}) + \varepsilon_i$, we assume

$$\mathbf{E}[\varepsilon_i|h(\boldsymbol{x}_i\,,\,\boldsymbol{\beta})] = 0\ , \quad i = 1, 2, \ldots, n$$

   i.e. disturbance at observation i is uncorrelated with the conditional mean function.

Assumptions relevant to the nonlinear regression model

4. **Homoscedasticity and non-autocorrelation:**

conditional homoscedasticity:

$$\mathbf{E}[\varepsilon_i^2 | h(\boldsymbol{x}_i\,,\, \boldsymbol{\beta})] = \sigma^2, \quad i = 1, 2, \ldots, n$$

non-autocorrelation:

$$\mathbf{E}[\varepsilon_t \varepsilon_s | h(\boldsymbol{x}_t\,,\, \boldsymbol{\beta}), h(\boldsymbol{x}_s\,,\, \boldsymbol{\beta})] = 0, \quad \text{for all } t \neq s$$

Assumptions relevant to the nonlinear regression model

5. **Data generating process:** DGP for $\boldsymbol{x}_i$ is assumed to be a well-behaved population such that first and second sample 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
6. **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**.

NLS: estimator of the nonlinear regression model

- NLS: min: $S(\boldsymbol{\beta}) = \sum[y_i - h(\boldsymbol{x}_i, \boldsymbol{\beta})]^2$
- Using standard procedure, we can get k first order conditions for the minimization:

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} [y_i - h(\boldsymbol{x}_i, \boldsymbol{\beta})] \frac{\partial h(\boldsymbol{x}_i \, , \, \boldsymbol{\beta})]}{\partial \boldsymbol{\beta}} = \mathbf{0}$$

- The above first order conditions are also moment conditions and this defines the NLS estimator as a GMM estimator.

NLS: estimator of the nonlinear regression model

- NLS being a GMM estimator allows us to deduce that the NLS estimator has good large sample properties: consistency and asymptotic normality (if assumptions are fulfilled).
- Hypothesis testing: The principal testing procedure is the Wald test, which relies on the consistency and asymptotic normality of the estimator. Likelihood ratio and LM tests can also be constructed.

Nonlinear regression: computing NLS estimates

For nonlinear models, a closed-form solution (NLS estimator) 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: examples

- LRM on TS with autocorrelation:

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + u_t \,, \qquad u_t = \rho u_{t-1} + \varepsilon_t,$$

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + \rho u_{t-1} + \varepsilon_t \qquad \text{note: } u_{t-1} = y_{t-1} - \boldsymbol{x}_{t-1}'\boldsymbol{\beta},$$

hence:

$$y_t = \rho y_{t-1} + \boldsymbol{x}_t'\boldsymbol{\beta} + \rho(\boldsymbol{x}_{t-1}'\boldsymbol{\beta}) + \varepsilon_t,$$

which is non-linear in parameters ($\rho\boldsymbol{\beta}$).

- Non-linear consumption function example:

$$cons_i = \beta_1 + \beta_2 inc_i^{\beta_3} + \varepsilon_i$$

special case: model is linear for $\beta_3 = 1$ (can be tested).

Examples 7.4 & 7.8 (Greene):
Analysis of a Nonlinear Consumption Function
OLS version: for $\beta_3 = 1$.

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: examples

Examples 7.4 & 7.8 (Greene):
Analysis of a Nonlinear Consumption Function
NLS with starting values equal to 0

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: examples

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

- Quantile regression estimates the relationship between regressors and a specified quantile of dependent variable.
- The (linear) quantile model can be defined as $Q[y|\boldsymbol{x}, q] = \boldsymbol{x}'\boldsymbol{\beta}_q$, such that $\text{Prob}[y \leq \boldsymbol{x}'\boldsymbol{\beta}_q|\boldsymbol{x}] = q$, $0 < q < 1$ where q denotes the q-th quantile of y.
- 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 ($q = \frac{1}{2}$).
- QREG (LAD) estimator can be motivated as a robust alternative to OLS (with respect to outliers).

Quantile regression (QREG)

For LRMs, the q-th quantile regression estimator $\boldsymbol{\beta}_q$ minimizes:

$$\min_{\hat{\boldsymbol{\beta}}_q}: \quad Q_n(\hat{\boldsymbol{\beta}}_q) = \sum_{i:\, e_i \geq 0}^{n} q|y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_q| \; + \sum_{i:\, e_i < 0}^{n} (1-q)|y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_q|,$$

where $e_i = (y_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}_q)$.

- We use the notation $\hat{\boldsymbol{\beta}}_q$ to make clear that different choices of q lead to different $\hat{\boldsymbol{\beta}}$.
- Slope of the loss function Q_n is asymmetrical (around $e_i = 0$).
- The loss function is not differentiable (at $e_i = 0$) $\rightarrow$ gradient methods are not applicable (linear programming can be used).

Quantile regression - LAD

- LAD estimator is the QREG for $q = \frac{1}{2}$ (median) and the loss function simplifies to:

$$\min_{\hat{\beta}_q}: \quad Q_n(\hat{\boldsymbol{\beta}}_q) = \sum_{i=1}^{n} |y_i - \boldsymbol{x}_i \hat{\boldsymbol{\beta}}_q|$$

- LAD estimator predates OLS (itself older than 200 years). Until recently, QREG and LAD have seen little use in econometrics, as OLS is vastly easier to compute.
- Different software packages use a variety of optimization algorithms for QREG/LAD estimation.
- Linear programming can be used for finding QREG estimates (Koenkerr and Bassett (around 1980).

Quantile regression

QREG coefficient interpretation example:

(1) $\text{wage}_i = \beta_1 + u_i$

(2) $\text{wage}_i = \beta_1 + \beta_2 \text{female}_i + u_i$

(3) $\text{wage}_i = \beta_1 + \beta_2 \text{female}_i + \beta_3 \, \text{exper}_i + u_i$

The above equations are estimated by OLS / LAD / QREG:

Coefficient	OLS	LAD ($q = \frac{1}{2}$)	QREG ($q = \frac{3}{4}$)
(1) β_1	$\hat{\beta}_1 = \overline{y}$ sample mean	$\hat{\beta}_1 = \tilde{y}$ sample median	$\hat{\beta}_1 = Q_3$ sample 3^{rd} quartile
(2) $\beta_1, \beta_1 + \beta_2$	conditional sample mean wage: male / female	cond. sample median wage: male / female	conditional sample Q_3 wage: male / female
(3) β_3	change in expected mean wage for $\Delta\text{exper} = 1$	change in exp. median wage for $\Delta\text{exper} = 1$	change in expected Q_3 wage for $\Delta\text{exper} = 1$

Example 7.10 (Greene):
Income Elasticity of Credit Cards Expenditure
OLS & LAD & Income elasticity at different 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			

Example 7.10 (Greene): Income Elasticity of Credit Cards Expenditure (LAD)

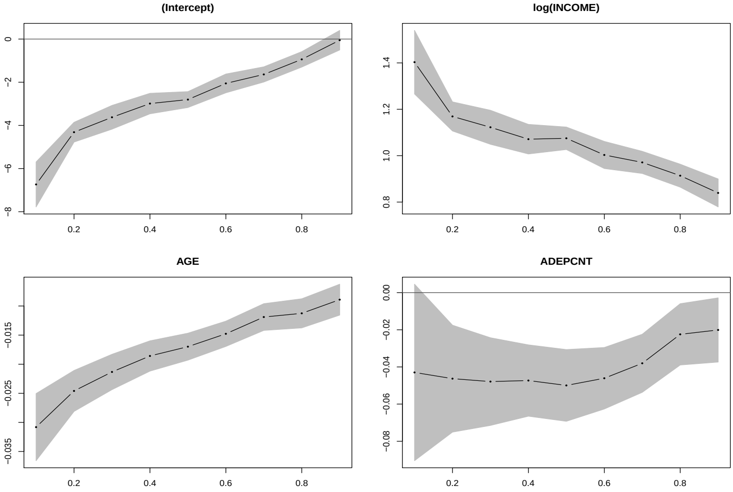
Depednent Variable: LOGSPEND   Method: Quantile Regression (Median)
Sample (adjusted): 3 13443   Included observations: 10499 after adjustments
Huber Sandwich Standard Errors & Covariance
Sparsity method: Kemel (Epanechnikov) using residuals
Bandwidth method: Hall-Sheather, bw=0.04437
Estimation successfully 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 example 2

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



Predictions from a model

- Predictions from a CLRM (repetition from BSc courses)
- Predictions: general features, kFCV, Variance vs. Bias

Predictions - basics

- CLRM and its estimate:

$$y = \beta_1 + \beta_2 x_3 + \beta_3 x_3 + \cdots + \beta_K x_K + u$$
$$\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \cdots + \hat{\beta}_K x_K$$

- Prediction of expected value:

$$\hat{y}_p = E(y|x_1 = 1, x_2 = c_2, \ldots, x_K = c_K)$$
$$\hat{y}_p = \hat{\beta}_1 + \hat{\beta}_2 c_2 + \hat{\beta}_3 c_3 + \cdots + \hat{\beta}_K c_K$$

- Rough (underestimated) confidence interval for the expected value prediction: (95%): $\hat{y}_p \pm 2 \times \text{s.e.}(\hat{y}_p)$. (Rule of thumb)

Predictions - basics

s.e.$(\hat{y}_p)$ can be obtained by reparametrization:

- Reparametrized CLRM:

$$y^* = \beta_1^* + \beta_2^*(x_2 - c_2) + \beta_3^*(x_3 - c_3) + \cdots + u$$

- The following holds:

$$\begin{aligned} \hat{y}_p &= \hat{\beta}_1^* \\ \text{s.e.}(\hat{y}_p) &= \text{s.e.}(\hat{\beta}_1^*), \quad i.e. \\ \text{var}(\hat{y}_p) &= \text{var}(\hat{\beta}_1^*) \end{aligned}$$

Predictions - basics

- Predicted and actual values of y_p:

$$\hat{y}_p = \hat{\beta}_1 + \hat{\beta}_2 c_2 + \hat{\beta}_3 c_3 + \cdots + \hat{\beta}_K c_K$$
$$y_p = \beta_1 + \beta_2 c_2 + \beta_3 c_3 + \cdots + \beta_K c_K + u_p$$

- Prediction error

$$\hat{e}_p = y_p - \hat{y}_p = (\beta_1 + \beta_2 c_2 + \beta_3 c_3 + \cdots + \beta_K c_K) + u_p - \hat{y}_p$$

- Prediction error variance

$$\mathrm{var}(\hat{e}_p) = \mathrm{var}(u_p) + \mathrm{var}(\hat{y}_p)$$

because $\mathrm{var}(\beta_1 + \beta_2 c_2 + \beta_3 c_3 + \cdots + \beta_K c_K) = 0$

Predictions - basics

- In CLRM, homoscedasticity holds, $\sigma^2 = \text{var}(u_p)$:
  - $\text{var}(\hat{e}_p) = \sigma^2 + \text{var}(\hat{y}_p)$
  - We estimate σ^2 from the original CLRM as $(SSR/(n-K-1))$
  - We get $\text{var}(\hat{y}_p)$ from the reparametrized LRM
- Standard prediction error:
  - $\text{s.e.}(\hat{e}_p) = \sqrt{\text{var}(\hat{e}_p)}$
- Prediction interval (95%)
  - $\hat{y}_p \pm t_{0.025} \times \text{s.e.}(\hat{e}_p)$

Predictions - basics

- Prediction with logarithmic dependent variable

$$\log(y) = \beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K + u$$
$$\widehat{\log(y)} = \hat{\beta}_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_K x_K$$

$\hat{y} = e^{\widehat{\log(y)}}$ systematically underestimates $\hat{y}$,

we can use a correction: $\hat{y} = \widehat{\alpha}_0 e^{\widehat{\log(y)}}$

where $\widehat{\alpha}_0 = n^{-1} \sum_{i=1}^{n} \exp(\hat{u}_i)$

is a consistent (not unbiased) estimator of $\exp(u)$.

Predictions - basics (Matrix form)

Prediction based on estimated model:

$$\hat{y}_p = \boldsymbol{x}_p' \hat{\boldsymbol{\beta}}$$

Difference between prediction and actual y_p value:

$$\hat{e}_p = \hat{y}_p - y_p = \boldsymbol{x}_p' \hat{\boldsymbol{\beta}} - \boldsymbol{x}_p' \boldsymbol{\beta} - u_p = \boldsymbol{x}_p' (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - u_p$$

If $\hat{\boldsymbol{\beta}}$ is unbiased estimator for $\boldsymbol{\beta}$,
$\hat{y}_p$ is an unbiased estimator for y_p value:

$$E(\hat{e}_p) = E(\hat{y}_p - y_p) = \boldsymbol{x}_p' E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + E(-u_p) = 0$$

and the variance of $\hat{e}_p$ can be expressed as:

$$E(\hat{e}_p^2) = \mathrm{var}(\hat{e}_p) = \boldsymbol{x}_p' \mathrm{var}(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_p + \mathrm{var}(u_p)$$

Predictions - basics (Matrix form)

Variance of $\hat{e}_p$ (continued):

$$\begin{aligned}
\mathrm{var}(\hat{e}_p) &= \boldsymbol{x}_p' \mathrm{var}(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_p + \mathrm{var}(u_p) \\
&= \boldsymbol{x}_p' \left[\sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right] \boldsymbol{x}_p + \mathrm{var}(u_p) \\
& \quad \text{substitute } \sigma^2, \mathrm{var}(u_p) \text{ with } \hat{\sigma}^2 \text{ (homoscedasticity)} \\
&= \underbrace{\boldsymbol{x}_p' \left[\hat{\sigma}^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right] \boldsymbol{x}_p}_{\hat{\sigma}_p^2} + \hat{\sigma}^2
\end{aligned}$$

With growing sample size (asymptotically),
$\mathrm{var}(u_p) = \hat{\sigma}_p^2 + \hat{\sigma}^2$ converges to $\hat{\sigma}^2$
$\ldots$ $\mathrm{plim}\,\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} \quad \leftrightarrow \quad \mathrm{plim}\,\hat{\sigma}_p^2 = 0$

Predictions - basics (Matrix form)

Variance of $\hat{e}_p$ (continued):

$$\text{var}(\hat{e}_p) \;=\; \boldsymbol{x}_p' \left[\hat{\sigma}^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right] \boldsymbol{x}_p + \hat{\sigma}^2$$

after re-arranging, s.e.$(\hat{e}_p)$ may be written as

$$\text{s.e.}(\hat{e}_p) \;=\; \hat{\sigma} \cdot \sqrt{1 + \boldsymbol{x}_p' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_p}\;,$$

which relates to the individual prediction error.

For mean prediction errors (considering $\hat{\sigma}_p^2$ only):

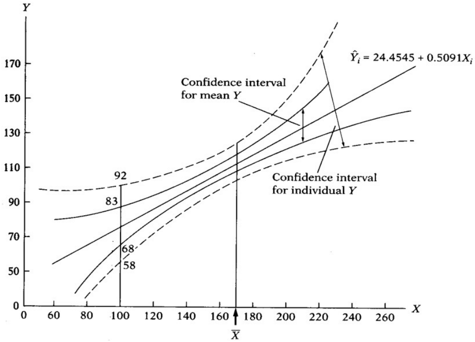
$$\text{s.e.}(\widetilde{e}_p) \;=\; \hat{\sigma} \cdot \sqrt{\quad \boldsymbol{x}_p' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_p}\;.$$

Predictions - basics (Matrix form)

Prediction intervals: individual vs. mean value predictions:

Individual prediction: $y_p \in \hat{y}_p \pm t^*_{\alpha/2} \times \text{s.e.}(\hat{e}_p)$

Mean value: $y_p \in \hat{y}_p \pm t^*_{\alpha/2} \times \text{s.e.}(\widetilde{e}_p)$



Predictions – general discussion:

- Reliability of predictions:
  - we work with estimated parameters
    (if we leave the CLRM paradigm, small sample properties of estimators may vary),
  - parameters can change in time (discussed separately in next Block – see Chow tests),
  - predictions include "individual" random errors.
- Impacts of random errors on predictions of individual values are usually much bigger than the impacts of variance in estimated parameters.

Mean Squared Error of prediction

We can generalize the previous discussion on predictions by considering both biased and unbiased predictors and by allowing for different functional forms and complexity levels in predictive models.

Predictions may be compared/evaluated using:

- $MSE = E\left[\left(y_i - \hat{f}(\boldsymbol{x}_i)\right)^2\right]$

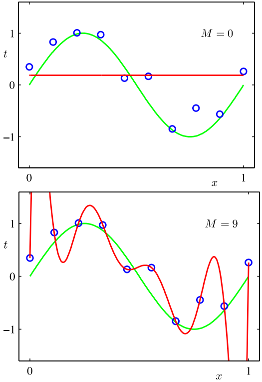
  where $\hat{f}(\boldsymbol{x}_i)$ is the prediction that $\hat{f}$ generates for the i-th regressor set. $\hat{f}$ represents a general class of predictors (linear, non-linear, non-parametric, etc.) and it may produce either biased or unbiased predictions

Variance vs. Bias trade-off

Population equation example: $y = \sin(x) + u$

Bias-Variance tradeoff – Intuition

- Model too simple: does not fit the data well
  - A *biased* solution
- Model too complex: small changes to the data, solution changes a lot
  - A *high-variance* solution



Train sample & Test sample

Suppose we fit a model $\hat{f}(\boldsymbol{x})$ to some training data $\mathrm{Tr} = \{y_i, \boldsymbol{x}_i\}_1^n$ and we wish to see how well it performs.

- We could compute *MSE* over Tr:

$$MSE_{\mathrm{Tr}} = \frac{1}{n} \sum_{i \in \mathrm{Tr}} \left[y_i - \hat{f}(\boldsymbol{x}_i) \right]^2$$

When searching for the "best" model by minimizing *MSE*, the above statistic would lead to over-fit models.

- Instead, we should (if possible) compute the *MSE* using fresh test data $\mathrm{Te} = \{y_i, \boldsymbol{x}_i\}_1^m$:

$$MSE_{\mathrm{Te}} = \frac{1}{m} \sum_{i \in \mathrm{Te}} \left[y_i - \hat{f}(\boldsymbol{x}_i) \right]^2$$

Variance vs. Bias trade-off

Suppose we have a model $\hat{f}(\boldsymbol{x})$, fitted to some training data Tr and let $\{y_0, \boldsymbol{x}_0\}$ be a test observation drawn from the population. If the true model is $y_i = f(\boldsymbol{x}_i) + \varepsilon_i$, with $f(\boldsymbol{x}_i) = \mathrm{E}(y_i|\boldsymbol{x}_i)$, then the **expected test MSE** can be decomposed into:

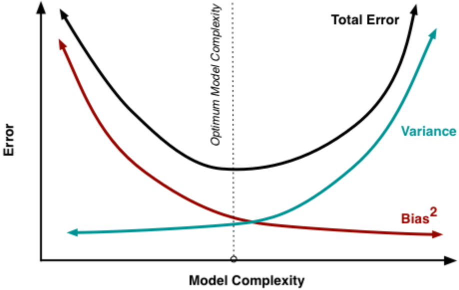
$E(MSE_0) = \mathrm{var}(\hat{f}(\boldsymbol{x}_0)) + [\mathrm{Bias}\,(\hat{f}(\boldsymbol{x}_0))]^2 + \mathrm{var}(\varepsilon_0)$,
where

$\mathrm{Bias}\,(\hat{f}(\boldsymbol{x}_0)) = E[\hat{f}(\boldsymbol{x}_0)] - f(\boldsymbol{x}_0)$,
ε_0 is the irreducible error: $E(MSE_0) \geq \varepsilon_0$,
all three RHS elements are non-negative,

The above equation refers to the average test MSE that we would obtain if we repeatedly estimated $f(\boldsymbol{x})$ using a large number of training sets and then tested each $\hat{f}(\boldsymbol{x})$ at $\boldsymbol{x}_0$.

Variance vs. Bias trade-off

$$E(MSE_0) = \mathrm{var}(\hat{f}(\boldsymbol{x}_0)) + [\mathrm{Bias}\,(\hat{f}(\boldsymbol{x}_0))]^2 + \mathrm{var}(\varepsilon_0),$$



This is an illustration, $\mathrm{var}(\varepsilon_0)$ not shown explicitly.
(lies at the /asymptotic/ minima of Variance and Bias^2)

k-Fold Cross Validation

- Training error (MSE_{Tr}) can be calculated easily.
- However, MSE_{Tr} is not a good approximation for the MSE_{Te} (out-of sample predictive properties of the model).
- Usually, MSE_{Tr} dramatically underestimates MSE_{Te}.

Cross-validation is based on re-sampling (similar to bootstrap).

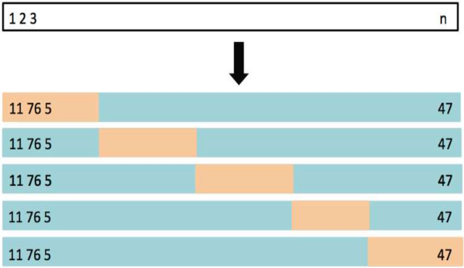
Repeatedly fit a model of interest to samples formed from the training set & make “test sample” predictions, in order to obtain additional information about predictive properties of the model.

k-Fold Cross Validation

- In k-Fold Cross-Validation (kFCV), the original sample is randomly partitioned into k roughly equal subsamples (divisibility).
- One of the k subsamples is retained as the test sample, and the remaining $(k-1)$ subsamples are used as training data.
- The cross-validation process is then repeated k times (the k folds), with each of the k subsamples used exactly once as the test sample.
- The k results from the folds can then be averaged to produce a single estimation.
- $k = 5$ or $k = 10$ is commonly used.

k-Fold Cross Validation

kFCV example for CS data & $k = 5$:
(random sampling, no replacement)



In TS, a similar "Walk forward" test procedure may be applied.

k-Fold Cross Validation

$$CV_{(k)} = \frac{1}{k} \sum_{s=1}^{k} MSE_s,$$

where $CV_{(k)}$ is the cross-validated estimate of MSE,
k is the number of folds used (e.g. 5 or 10),
$MSE_s = \frac{1}{m_s} \sum_{i \in C_s} (y_i - \widehat{y}_i)^2$
m_s is the number of observations in the s-th test sample
C_s refers to the s-th set of test sample observations.

As we evaluate predictions from two or more models,
we look for the lowest $CV_{(k)}$.