Testing the Coherence of Data and External Intervals

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Martin Jann
martin.jann@uni-hamburg.de

Background

Task: Test whether data originated from specific population given a model and external information.

Problem: External information is **sparse** (only moment type information is known) and **uncertain** (only bounds on the true value are known).

Techniques for incorporating external information: Prior distributions in Bayesian statistics, Constrained optimization, Generalized Method of Moments (GMM) [2]
$\Rightarrow$ Using GMM, external moment type information can be directly combined with moment conditions used to estimate a statistical model [3], and the **Sargan-Hansen test** [2] becomes a coherence test.

Scope of GMM: OLS, Maximum Likelihood (score function), M-estimation and many more

Example

Simple linear regression model: $y = \beta_1 + x\beta_2 + \epsilon$

External information: $E(y) = 100$

Assuming $E(\epsilon) = 0$, the value $E(y) = 100$ becomes a constraint on the parameter,

$$100 = E(y) = \beta_1 + E(x)\beta_2. \quad (1)$$

How can (1) be used in constrained optimization or to identify possible prior distributions?

Point-Valued Case

Construction of the **Sargan-Hansen test** for n given data vectors $\mathbf{z}_i$ i.i.d like $\mathbf{z}$, model parameters $\boldsymbol{\theta}$ and external (point) information $\mathbf{e}$:

1. Define moment functions $\mathbf{m}(\mathbf{z},\boldsymbol{\theta})$ (for the model) and $\mathbf{h}(\mathbf{z},\mathbf{e})$ (for the external information), both with expected value $\mathbf{0}$, and let $\mathbf{g}(\mathbf{z},\boldsymbol{\theta}) := (\mathbf{m}(\mathbf{z},\boldsymbol{\theta})^T, \mathbf{h}(\mathbf{z})^T)^T$.
2. Find a consistent, (almost certainly) positive-definite estimator, $\hat{\boldsymbol{\Omega}}$ of $E(\mathbf{g}(\mathbf{z},\boldsymbol{\theta})\mathbf{g}(\mathbf{z},\boldsymbol{\theta})^T)$ and let $\hat{\mathbf{W}} := \hat{\boldsymbol{\Omega}}^{-1}$, if the inverse exists.
3. Determine the maximum value $\hat{Q}$ of the objective function $\hat{Q}_n(\boldsymbol{\theta}) = -(\frac{1}{n}\sum_{i=1}^n \mathbf{g}(\mathbf{z}_i,\boldsymbol{\theta}))^T \hat{\mathbf{W}} (\frac{1}{n}\sum_{i=1}^n \mathbf{g}(\mathbf{z}_i,\boldsymbol{\theta}))$.

$\Rightarrow$ Under the regularity conditions for GMM $-n\hat{Q} \xrightarrow{d} \chi^2_{p-q}$.

Main result:

$$\overline{\mathbf{m}} - \widehat{\mathbf{Cov}}(\mathbf{h},\mathbf{m})^T \widehat{\mathbf{Var}}(\mathbf{h})^{-1}\overline{\mathbf{h}} = \mathbf{0} \text{ for a } \boldsymbol{\theta}_h \Rightarrow -n\hat{Q} = n\overline{\mathbf{h}}^T \widehat{\mathbf{Var}}(\mathbf{h})^{-1}\overline{\mathbf{h}}$$

$\Rightarrow$ **In many relevant cases (e.g. OLS), the test statistic is a function of the external information and the data alone.**

Chosen examples of $\widehat{\mathbf{Var}}(\mathbf{h})$:

- $\hat{\boldsymbol{\Sigma}}_h = \frac{1}{n}\sum_{i=1}^n \mathbf{h}(\mathbf{z}_i)\mathbf{h}(\mathbf{z}_i)^T$ (dependent on $\mathbf{e}$)
- $\hat{\mathbf{S}}_h = \frac{1}{n-1}\sum_{i=1}^n (\mathbf{h}(\mathbf{z}_i) - \overline{\mathbf{h}})(\mathbf{h}(\mathbf{z}_i) - \overline{\mathbf{h}})^T$ (independent of $\mathbf{e}$)

Interval-Valued Case

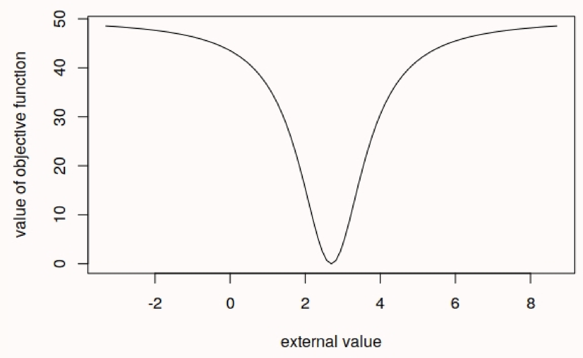
A closed interval $\mathbf{I}_{ex}$ containing the true external value $\mathbf{e}_0$ is known.
$\Rightarrow$ Interval of possible test statistics $[\underline{\chi}^2, \overline{\chi^2}]$ (**independent of $\boldsymbol{\theta}$!**) with asymptotic $\chi^2(\lambda)-$distributions, where $\lambda \in [0,\infty)$.

Desired property of a coherence test:
Test decides that $\mathbf{e}_0 \notin \mathbf{I}_{ex}$. $\Rightarrow$ Test decides that $\mathbf{e}_0 \notin \mathbf{I}$ for all $\mathbf{I} \subset \mathbf{I}_{ex}$.

$\Rightarrow$ **$\Gamma-$maximin rule for selecting a p-value**

- Credal set: Indicator functions of all events $\{\chi^2_{\mathbf{e}} \leq -n\hat{Q}\}$, where $\chi^2_{\mathbf{e}} \in [\underline{\chi}^2, \overline{\chi^2}]$ is fixed in each case (p-value events)
- Lower probabilities are obtained for the central χ^2-distribution, because $\chi^2(\lambda)-$distributions are **stochastically ordered in λ**.
- The maximum lower probability (p-value of the interval test) is obtained with $\underline{\chi}^2$ and a central χ^2-distribution.

Results: Two asymptotic tests, $\mathbf{SH}(\hat{\mathbf{S}}_h)$ and $\mathbf{SH}(\hat{\boldsymbol{\Sigma}}_h)$; Small sample tests under the normality assumption are the same for $\hat{\boldsymbol{\Sigma}}_h$ and $\hat{\mathbf{S}}_h$ $\Rightarrow$ **IDC**.



Simulations

Scenario: $x \sim N(4,4)$ and $y = 16 + x + \epsilon$, with $\epsilon \sim N(0,60)$ $\Rightarrow$ linear regression under Gauss-Markov assumptions with $R^2 = 0.0625$

Sampling: (x_i, y_i) i.i.d like (x,y) with $i = 1,\ldots,n$ for $n = 30$ and $n = 50$

External moments: $E(x)$, $E(y)$, $\text{Var}(y)$ (using the sample variance formula) and combinations thereof

$\mathbf{I}_{ex}$: Let $\mathbf{e}_0$ be the true moment value $\Rightarrow [0.95\cdot\mathbf{e}_0, 1.05\cdot\mathbf{e}_0]$ (type 1 error setting) and $[1.2\cdot\mathbf{e}_0, 1.3\cdot\mathbf{e}_0]$ (power setting)

For each of the 28 scenarios, the null hypothesis rejection rates of the three tests were calculated in R ($\alpha = 0.05$), simulating each scenario 10000 times.

Results

- Type I errors of $E(x)$ and $E(y)$ are substantially smaller than α and combinations of moments do not lead to a substantial increase in power; **Conservativity: $\mathbf{SH}(\hat{\mathbf{S}}_h) <$ IDC $< \mathbf{SH}(\hat{\boldsymbol{\Sigma}}_h)$**
- Using $\text{Var}(y)$ alone leads to high type I error rates and low power. $\Rightarrow$ The true distribution deviates too much from the normal distribution.
- Using $\text{Var}(y)$ in combination with other moments can reduce type I error rates and increase power compared to using $\text{Var}(y)$ alone.

Further Research

- Use the $\Gamma-$maximin decision rule to introduce (external) intervals into econometric and psychometric procedures
- Compare interval-valued GMM approach with generalized Bayes approach
- Analyze the level of robustness

References

[1] Thomas Augustin, Gero Walter and Frank P. A. Coolen. Statistical inference. In *Introduction to Imprecise Probabilities*,, John Wiley & Sons, Ltd, 2014.

[2] Lars P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4): 1029–1054, 1982.

[3] Guido W. Imbens and Tony Lancaster. Combining micro and macro data in microeconometric models. *The Review of Economic Studies*, 61(4):655–680, 1994.