Dwight Look College of ENGINEERING
TEXAS A&M UNIVERSITY

Insights from Machine Learning for Evaluating Production Function Estimators on Manufacturing Survey Data

José Luis Preciado Arreola and Andrew L. Johnson

Associate Professor, Department of Industrial and Systems Engineering

Visiting Associate Professor, School of Information Science and Technology

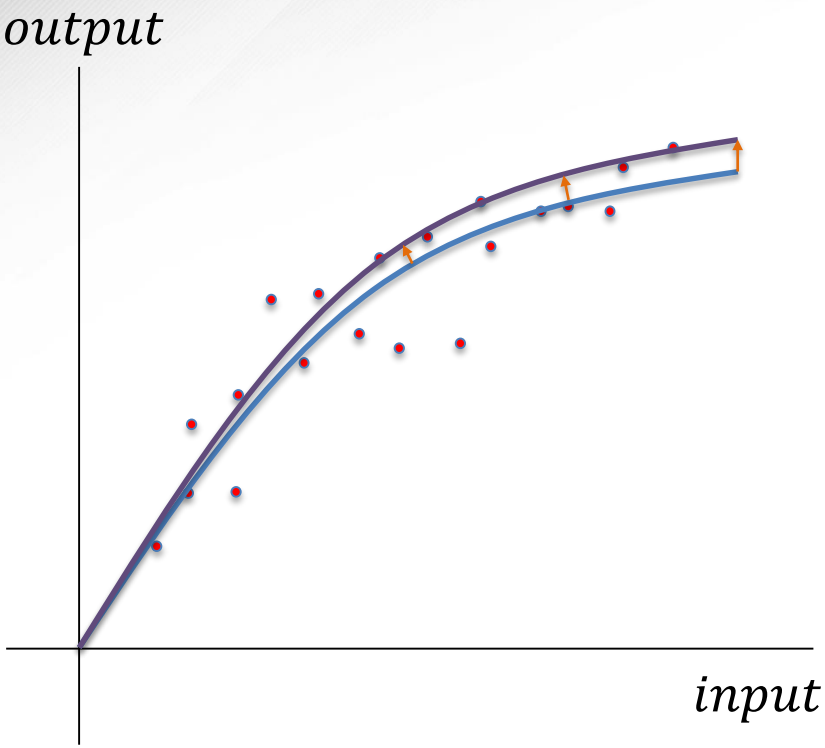
Government Data

7th largest country in South American in terms of Area

6th largest country in South American in terms of Population

- Issues Related to Census Manufacturing Data
  - Estimator Development
  - Which Estimator to Use?
  - If a Census is not available, how large of a survey is needed?
- General Framework to Evaluate Production Functions
- Analysis of Chilean data

Stochastic Frontier Model



$$y_i = f(\boldsymbol{x}_i) + v_i - u_i$$
$$i = 1, \dots, n$$

$f(\cdot)$ (***Production frontier***):

- Monotonic and concave (DRS).
- $\mathbb{R}^d \longrightarrow \mathbb{R}$

v_i (***Noise***)

- Mean zero: $E(v_i) = 0$

u_i (***Inefficiency***)

- Non-negative: $u_i \geq 0, \ i = 1, \dots, n$

Estimators

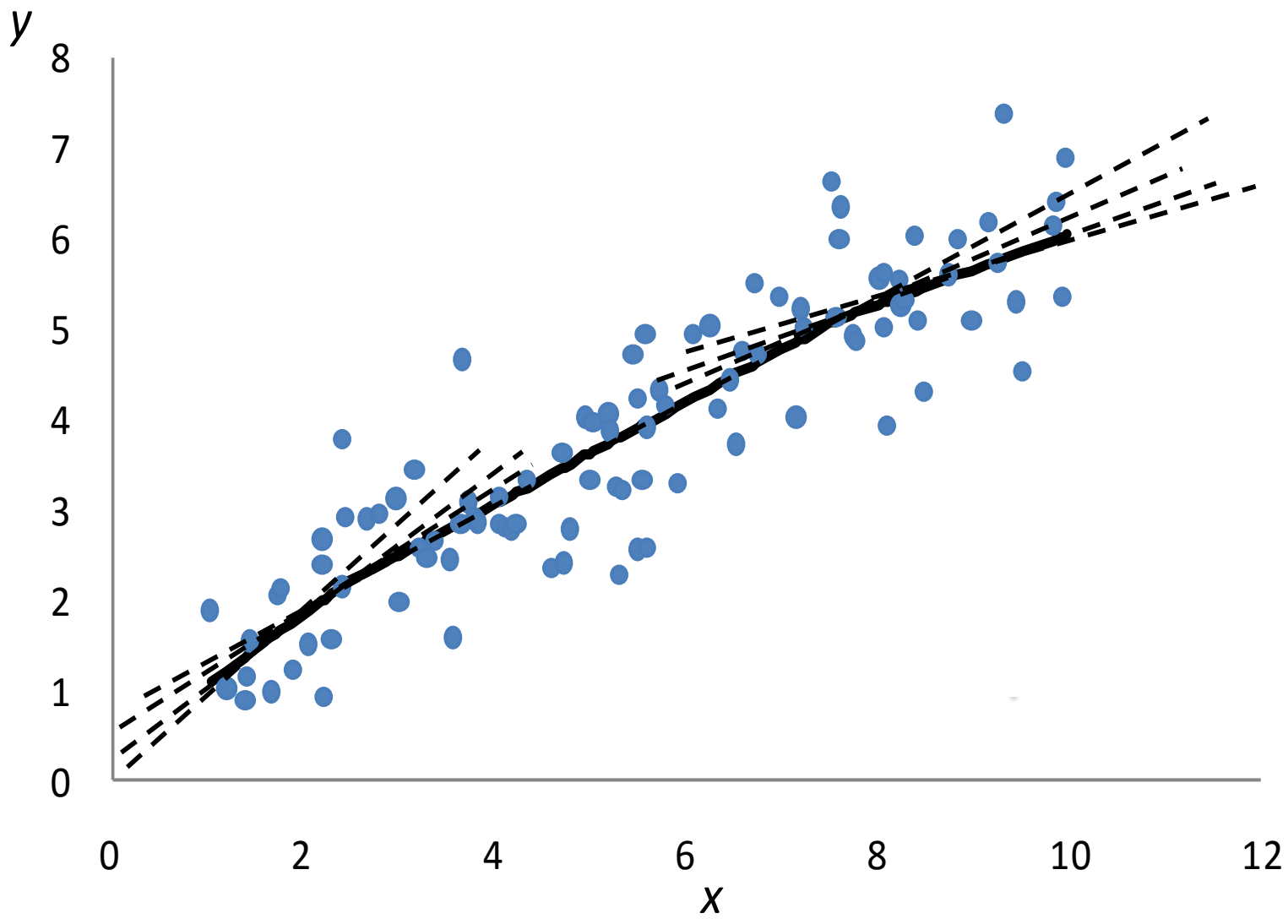
There is a large variety of methods to estimate production functions.

Functional Estimator	One Stage/ Two Stage	Noise	Parametric/ Nonparametric	Flexible u_i Variance	Scalability
DEA	1 S	No	NP	Yes	Large
ML SFA	1 S	Yes	P	Possible	Large
Bayesian SFA	1 S	Yes	P	Possible	Low 10000's
Banker-Maindiratta	1 S	Yes	SP	No	Low 100's
KPST	1 S	Yes	NP	Yes	Low 100's
MBCR-I	1 S	Yes	SP	Limited	Mid 1000's
OLS (CD, TL, CES)	2 S	Yes	P	No	Large
CNLS	2 S	Yes	NP	No	Mid 1000's
CWB	2 S	Yes	NP	No	Mid 100's
CAP	2 S	Yes	NP	No	Low 10000's
SCKLS	2 S	Yes	NP	No	Gridded

Production Theory

- Monotonicity
  - More input should be able to generate more output
- Concavity (of output in input)
  - Decreasing marginal benefits of additional inputs
- Convex input sets
  - Substitutability between capital and labor with some optimal mix

Estimators



Convex Nonparametric Least Squares (CNLS)

$$\min_{\alpha,\beta,\varepsilon} \left\{ \sum_{i=1}^{n} \varepsilon_i^2 \;\middle|\; \begin{aligned} & y_i = \alpha_i + \beta_i^{\mathrm{T}} \mathbf{x}_i + \varepsilon_i \quad \text{for } i = 1, \ldots, n && (1a) \\ & \alpha_i + \beta_i^{\mathrm{T}} \mathbf{x}_i \leq \alpha_h + \beta_h^{\mathrm{T}} \mathbf{x}_i \quad \text{for } i, h = 1, \ldots, n \text{ and } i \neq h && (1b) \\ & \beta_i \geq 0 \quad \text{for } i = 1, \ldots, n, && (1c) \end{aligned} \right\}$$

- 1st constraint: linear regression
- 2nd constraint: convexity using Afriat inequalities
- 3rd constraint: monotonicity

– Computation burden
  - 2nd constraints will generate n(n-1) constraints, where n is number of observations

Estimators

- Convex adaptive partitioning CAP (Hannah and Dunson, 2013),

$$\hat{f}(\boldsymbol{x}_i) = \min_{k \in \{1,\ldots,K\}} \beta_{0[i]} + \boldsymbol{\beta}^{\boldsymbol{T}}_{-\mathbf{0}[i]} \boldsymbol{x}_i$$

$$\boldsymbol{\beta}_{-0\boldsymbol{k}} \geq \mathbf{0}, \forall\, k = 1, \ldots K$$

+

Greedy Partitioning Algorithm to determine $[i]$ mappings

CAP

Starts with simple model ($K = 1$ hyperplane)

One-to-many hyperplane-observation mapping

Series of OLS regression problems

CAP-NLS formulation and details

$$\hat{f}_K(\boldsymbol{X}_i) = \beta^*_{0[i]} + \boldsymbol{\beta}^{*T}_{-0[i]}\boldsymbol{X}_i$$

$$(\beta^*_{0k}, \boldsymbol{\beta}^*_{-0k})_{i=1}^K = \underset{(\beta_{0k}, \boldsymbol{\beta}_{-0k})_{i=1}^K}{argmin} \sum_{i=1}^n \epsilon_i^2$$

$$s.t. \quad \epsilon_i = \beta_{0[i]} + \boldsymbol{\beta}^T_{-0[i]}\boldsymbol{X}_i - Y_i$$

$$\beta_{0[i]} + \boldsymbol{\beta}^T_{-0[i]}\boldsymbol{X}_i \leq \beta_{0k} + \boldsymbol{\beta}^T_{-0k}\boldsymbol{X}_i$$

$$\forall\, i = 1, \dots n, k = 1, \dots, K$$

$$\boldsymbol{\beta_{-0k}} \geq \mathbf{0}, \forall\, k = 1, \dots K$$

+

Greedy Partitioning Algorithm to determine $[i]$ mappings

CAP-NLS

Starts with simple model ($K = 1$ hyperplane)

One-to-many hyperplane-observation mapping

Conditionally Global optimization (QCP)

Can be formulated as series of quadratic programs

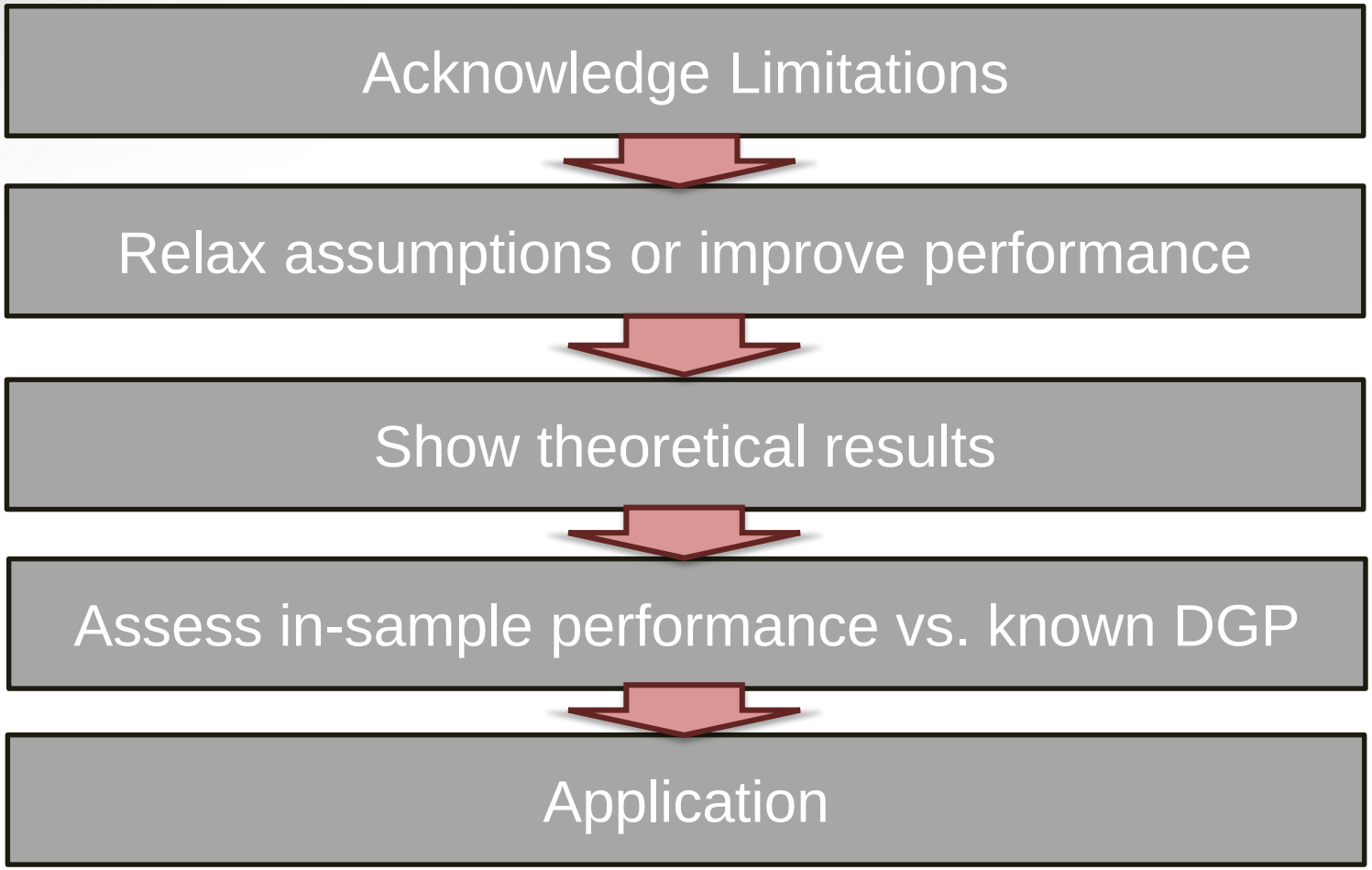
CAP-NLS is …

	CAP	CNLS	CAP-NLS
Hyperplane fitting	Local, Myopic, Refit-corrected	Global	Global
Partitioning strategy	Greedy Adaptive	"Big" Model	Greedy Adaptive
Concavity imposing	Min. of Hyperplanes	Afriat Inequalities	Afriat Inequalities

- Globally optimizes given an observation-hyperplane assignment
- Explores observation-hyperplane assignments adaptively

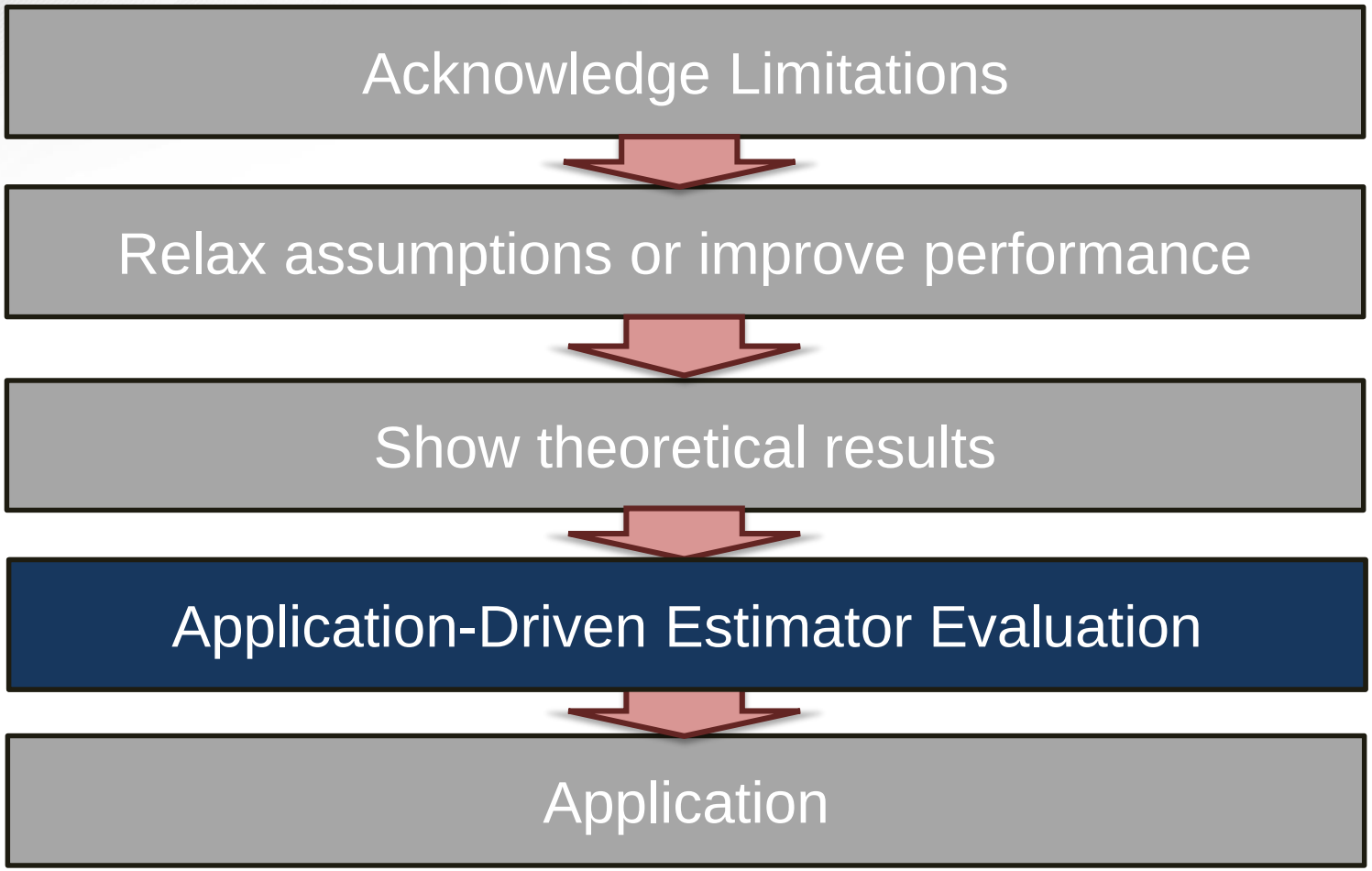
Application-Driven Estimator Evaluation

*How are functional estimators being **proposed/compared**?*



Application-Driven Estimator Evaluation

*How are functional estimators being **proposed/compared**?*



Application-Driven Estimator Evaluation

Bridging the gap…

How would we estimate these errors on simulated datasets?

Define

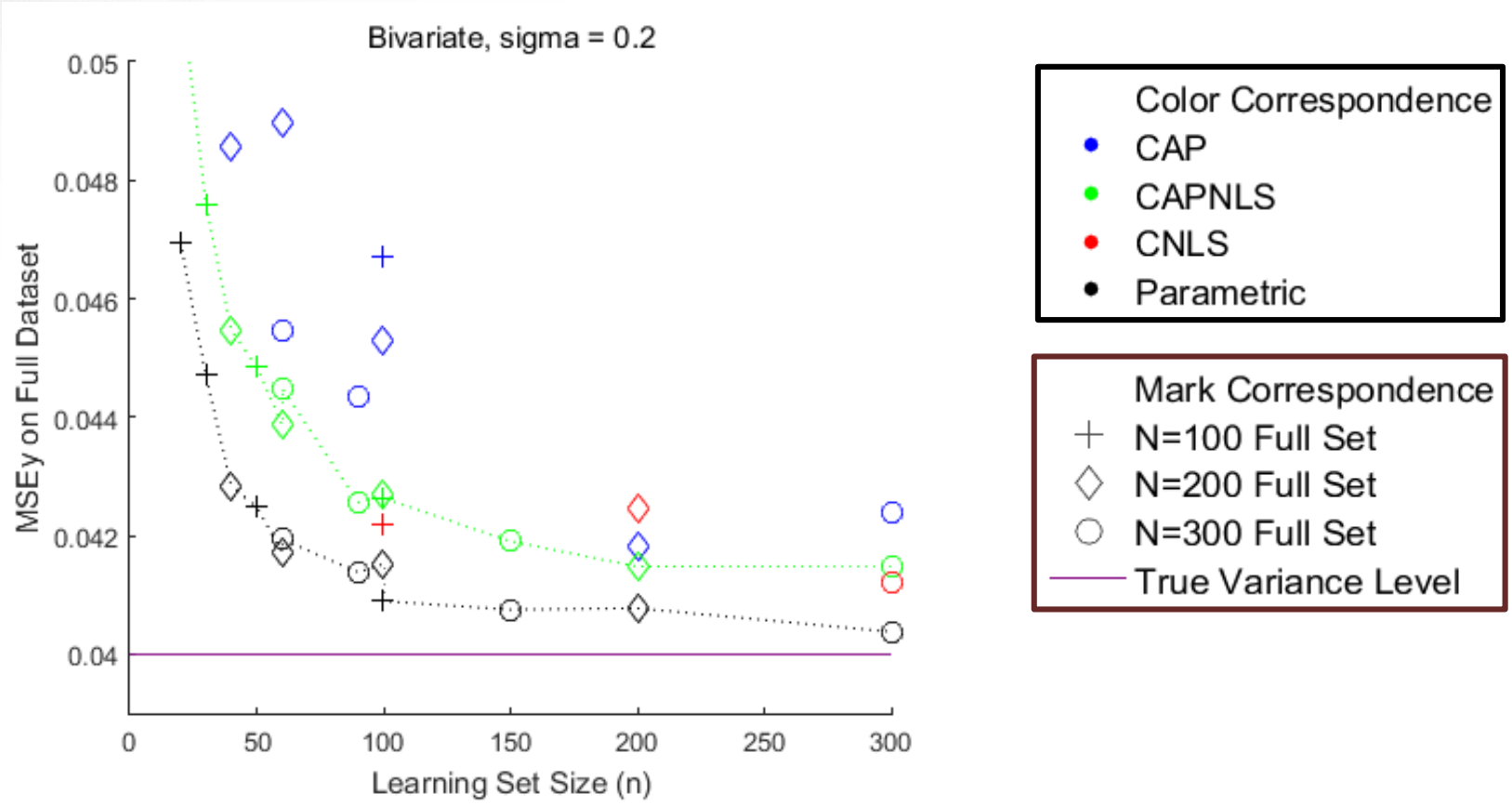
$E(\widehat{Err_{ISy}}) = \overline{MSE_{yIS}} =$ Test against many data sets with same input vector, but different error vectors.

$E(\widehat{Err_{Pf}}) = \overline{MSE_{Pf}} =$ Test against many large, unobserved datasets with different input and error vectors.

$E(\widehat{Err_{FS.}}) = \overline{MSE_{FS}}. =$ Weight in-sample and predictive errors

Simulated Data Results

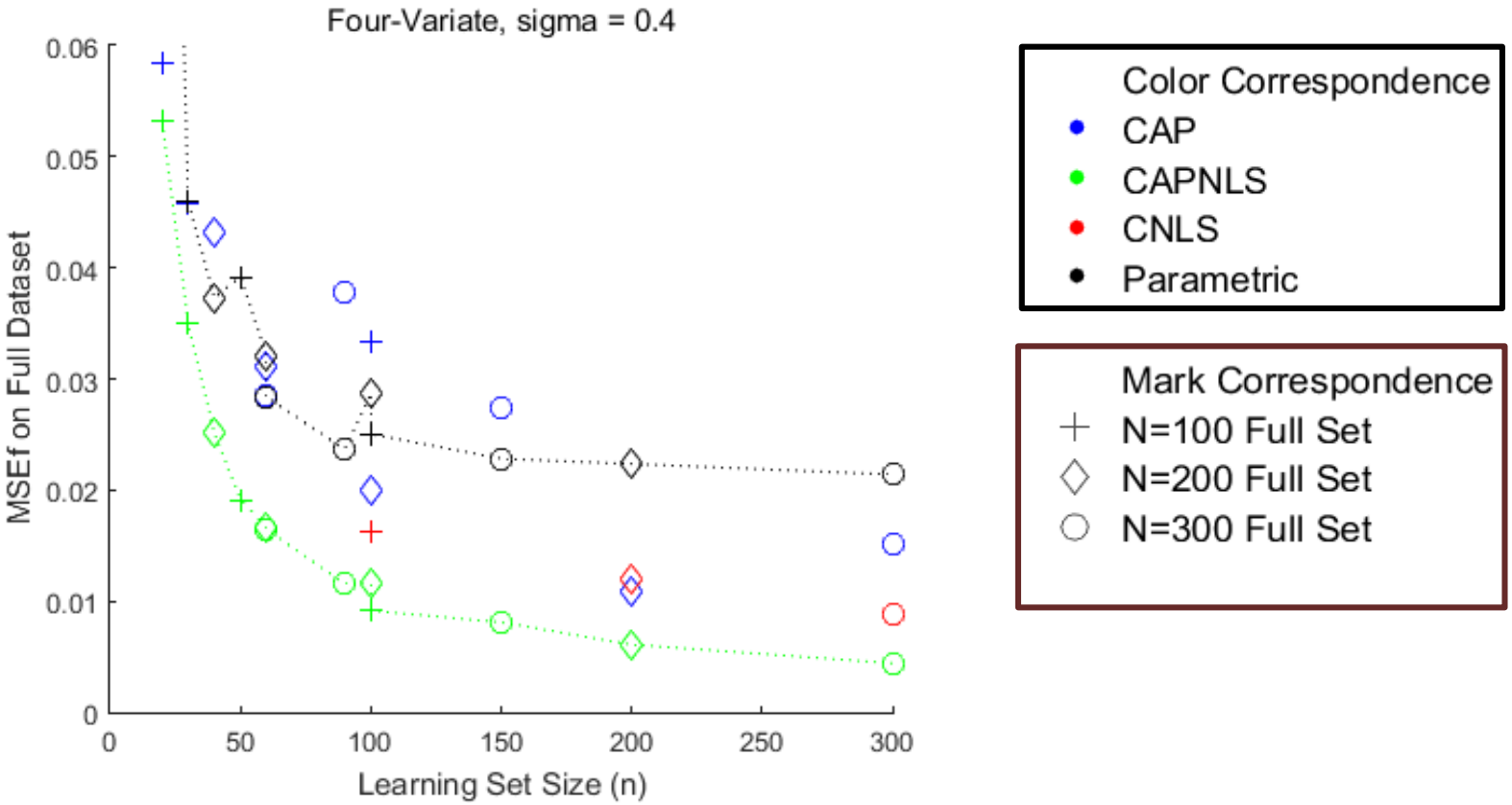
$$Y_i = X_{i1}^{0.4} X_{i2}^{0.5} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \sigma = 0.2, X_{ij} \sim Unif(.1,1) \; \forall i, j$$



Simulated Data Results

How about we double the number of inputs?

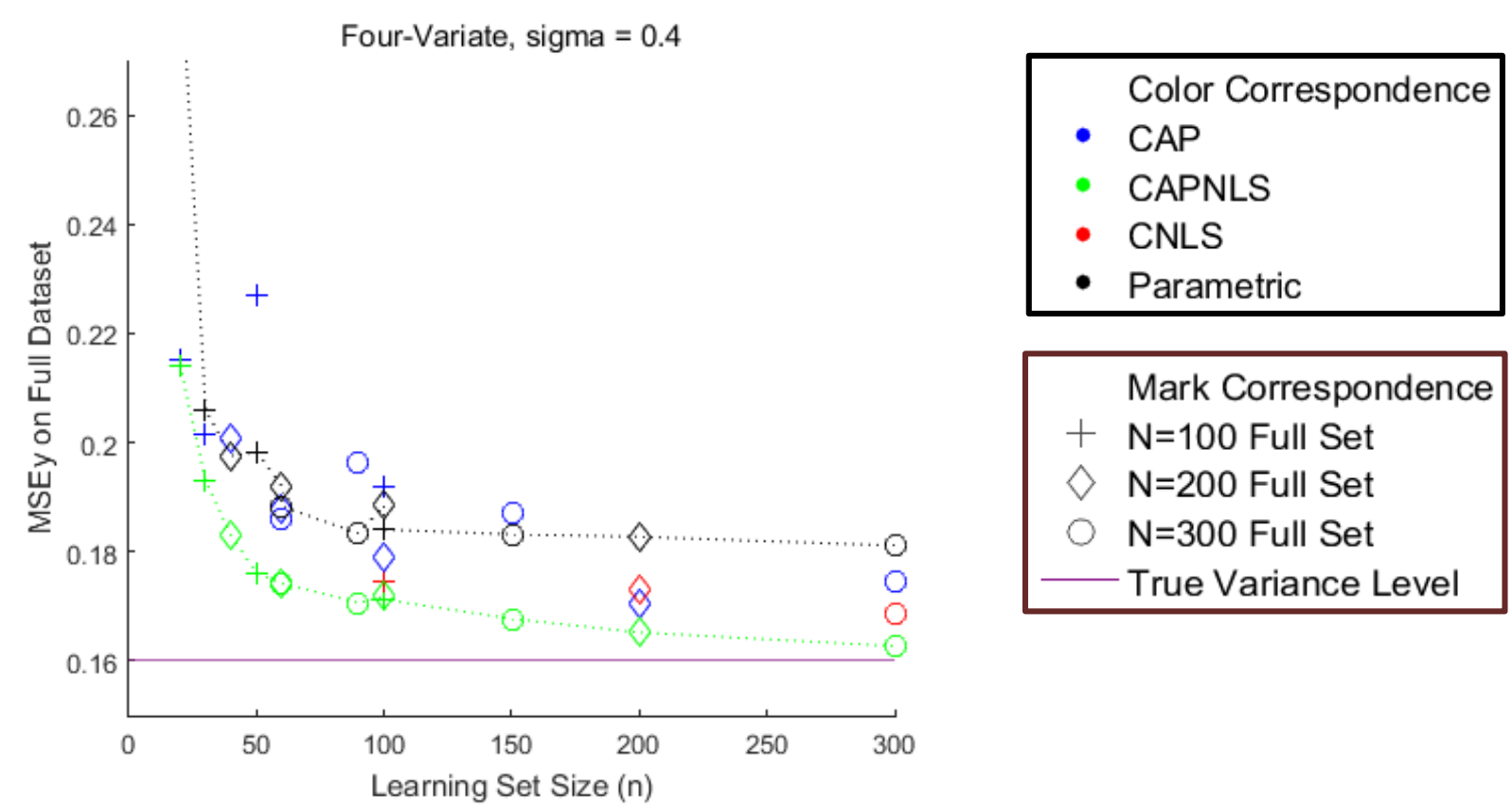
$$Y_i = X_{i1}^{0.3} X_{i2}^{0.25} X_{i3}^{0.25} X_{i4}^{0.1} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)\ , \sigma = 0.2, X_{ij} \sim Unif(.1,1)\ \forall i, j$$



Simulated Data Results

How about we double the number of inputs?

$$Y_i = X_{i1}^{0.3} X_{i2}^{0.25} X_{i3}^{0.25} X_{i4}^{0.1} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \sigma = 0.2, X_{ij} \sim Unif(.1,1) \; \forall i, j$$



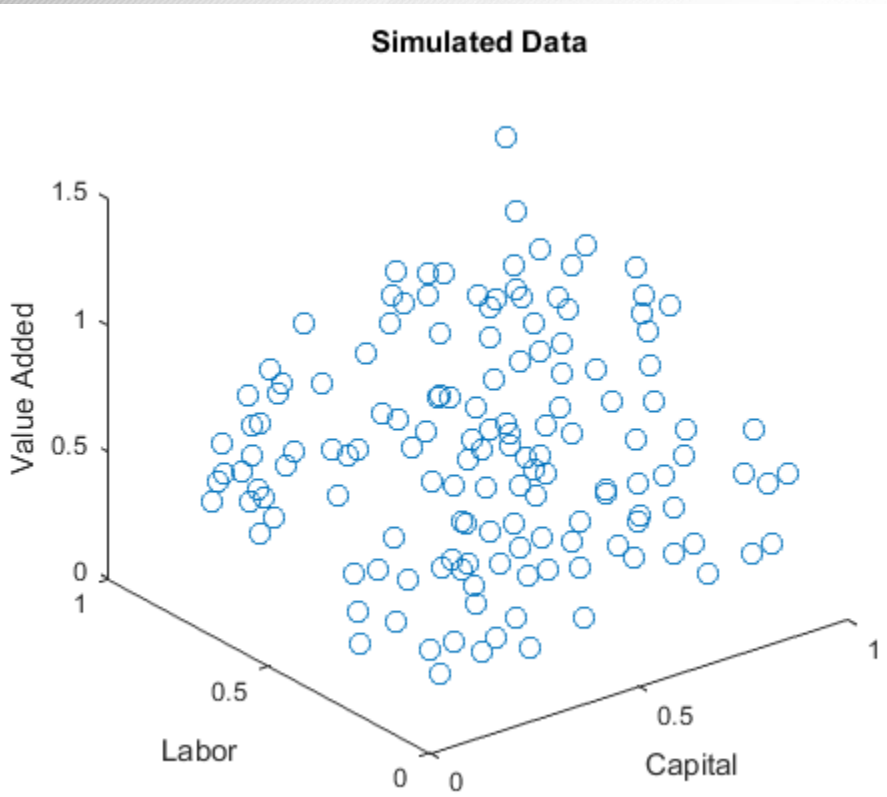
Simulated Data Results

		CAP-NLS			CAP			CNLS		
σ (noise std. dev.)	nL (learning set size)	100	200	300	100	200	300	100	200	300
0.2	K	7	12	13	2	2	2	63	132	192
	Time (s)	4	36	79	0.45	1	2	1	9	31
0.3	K	7	12	12	2	2	2	59	122	192
	Time (s)	4	36	74	0.46	1	2	1	10	32
0.4	K	7	12	12	2	2	2	57	122	186
	Time (s)	4	36	75	0.48	1	2	1	10	33

CAP-NLS does best both on the “traditional” testing framework and in the census (full set)-fit, application-driven one.

Let’s use the A-D evaluation framework on **real** data now.

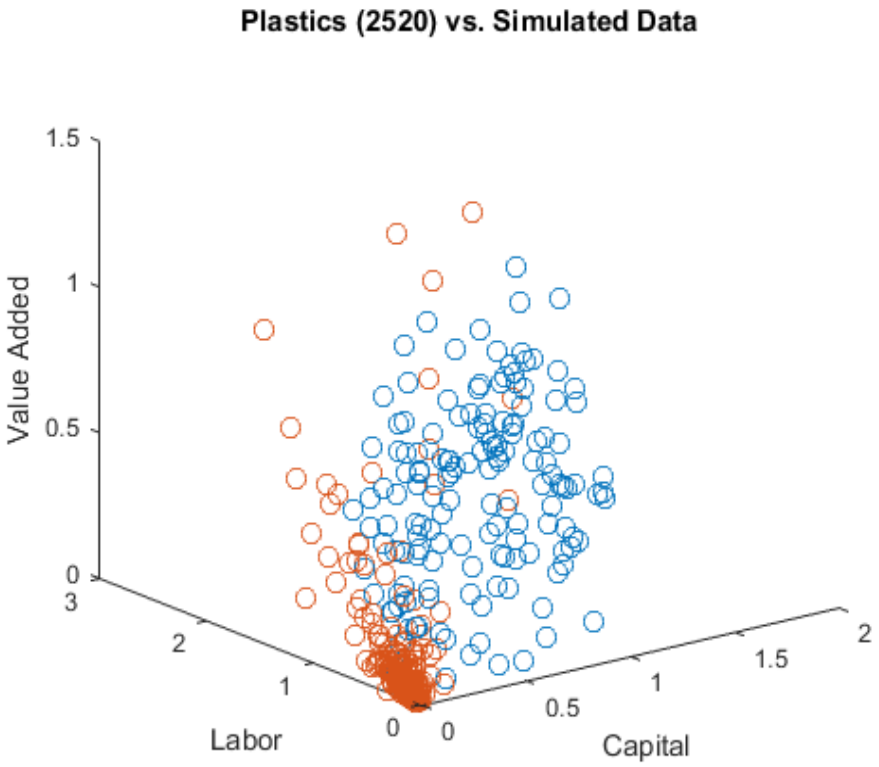
Chilean Manufacturing



- SFA literature has focused on illustrating performance on simulated data (after demonstrating consistency results) and extrapolating.

Chilean Manufacturing

Using Chilean Manufacturing Survey Data (ENIA, 2010)



- Data clustered at popular scale sizes
- Data clustered at popular input ratios.
- Input-specific marginal distributions.
- Some large firms have **different** input ratios.

Chilean Manufacturing

Can only test against y now..

We do not have the DGP anymore, thus define **different estimators** for error quantities.

Error		Error Estimator
ERR_{ISy} ←	In-Sample Error	Parametric Bootstrap (Efron, 2004)
ERR_{Py} ←	Predictive Error	Repeated Learning Testing (Breiman et al., 1984)
ERR_{FSy} ←	Finite Full Census Error	Weighting

Chilean Manufacturing

Using Chilean Manufacturing Survey Data (ENIA, 2010)

Industry Name and Code	n	Survey Size	R^2_{FS}	K_{nL}^{CAPNLS}	Best Method
Other Metal Products (2899)	144	20%	50%	1	CAP-NLS, CDA
		30%	60%	2	CAP-NLS, CDA
		40%	64%	2	CAP-NLS, CDA
		50%	72%	3	CAP-NLS
		100%	88%	7	CAP-NLS
Wood (2010)	150	20%	35%	1	CDA
		30%	40%	1	CAP-NLS, CDA
		40%	47%	2	CAP-NLS, CDA
		50%	52%	3	CAP-NLS, CDA
		100%	66%	6	CAP-NLS
Structural Use Metal (2811)	161	20%	77%	1	CAP-NLS, CAP
		30%	82%	2	CAP-NLS
		40%	87%	3	CAP-NLS, CAP
		50%	90%	4	CAP-NLS
		100%	95%	9	CAP-NLS, CAP

Chilean Manufacturing

Using Chilean Manufacturing Survey Data (ENIA, 2010)

Industry Name and Code	n	Survey Size	R^2_{FS}	K_{nL}^{CAPNLS}	Best Method
Plastics (2520)	249	20%	54%	2	CAP-NLS, CAP, CDA
		30%	57%	3	CDA
		40%	57%	5	CAP-NLS, CAP, CDA
		50%	60%	7	CAP-NLS, CAP, CDA
		100%	64%	11	CAP-NLS, CAP, CDA
Bakeries (1541)	250	20%	72%	3	CAP
		30%	77%	3	CAP
		40%	78%	4	CAP, CDA
		50%	85%	4	CAP
		100%	99%	5	CAP-NLS, CAP, CDA

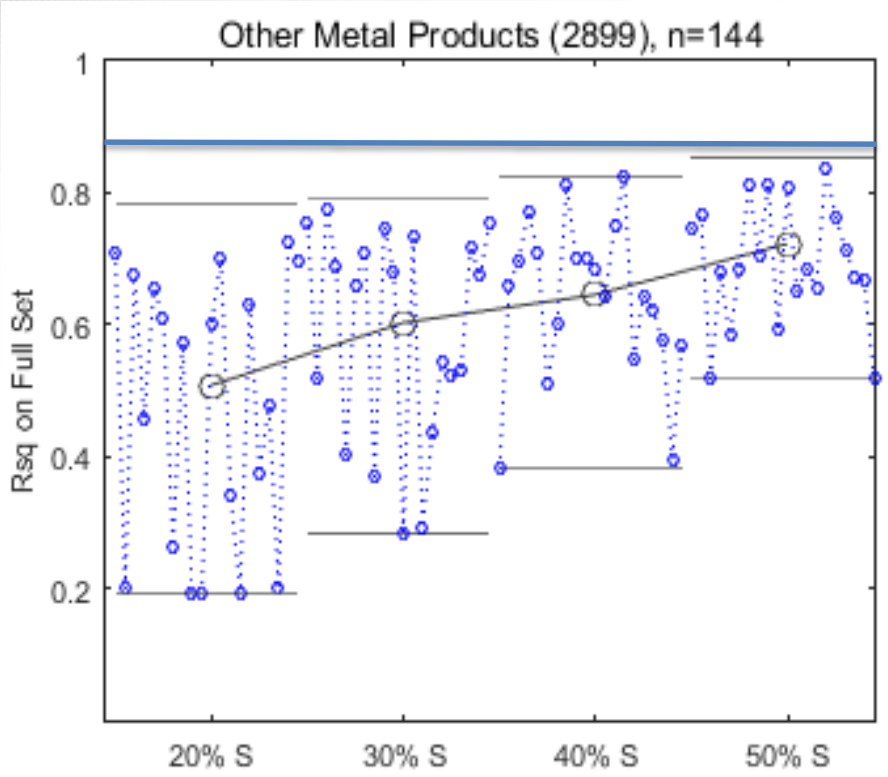
Chilean Manufacturing

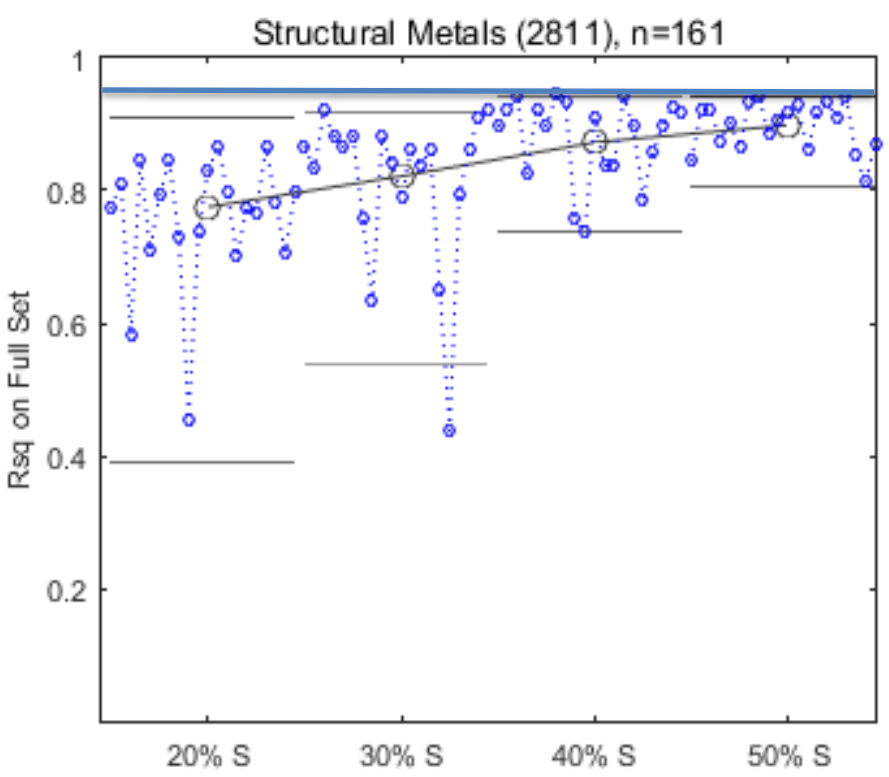
How well does Cobb-Douglas fit?

Industry Name and Code	n	Survey Size	R^2_{FS}	R^2_{CDA}	Ratio vs. Best Method
Other Metal Products (2899)	144	20%	50%	49%	CDA ties for Best Method
		30%	60%	59%	CDA ties for Best Method
		40%	64%	64%	CDA ties for Best Method
		50%	72%	60%	0.83 vs. CAP-NLS
		100%	88%	79%	0.90 vs. CAP-NLS
Wood (2010)	150	20%	35%	35%	CDA ties for Best Method
		30%	40%	40%	CDA ties for Best Method
		40%	47%	47%	CDA ties for Best Method
		50%	52%	51%	CDA ties for Best Method
		100%	66%	62%	0.94 vs. CAP-NLS
Structural Use Metal (2811)	161	20%	77%	69%	0.90 vs. CAP-NLS
		30%	82%	76%	0.93 vs. CAP-NLS
		40%	87%	81%	0.93 vs. CAP-NLS
		50%	90%	87%	0.97 vs. CAP-NLS
		100%	95%	91%	0.96 vs. CAP-NLS

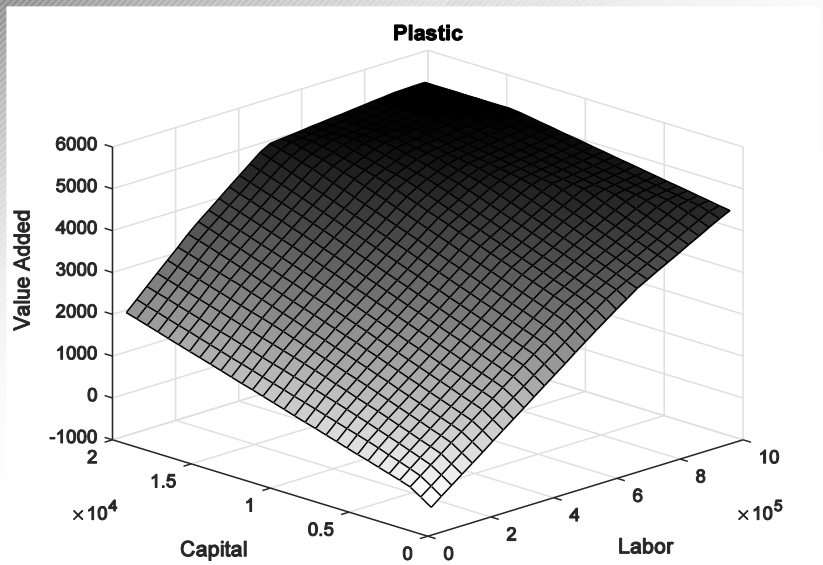
Chilean Manufacturing

How good/reliable is a production function from survey data?





Chilean Manufacturing



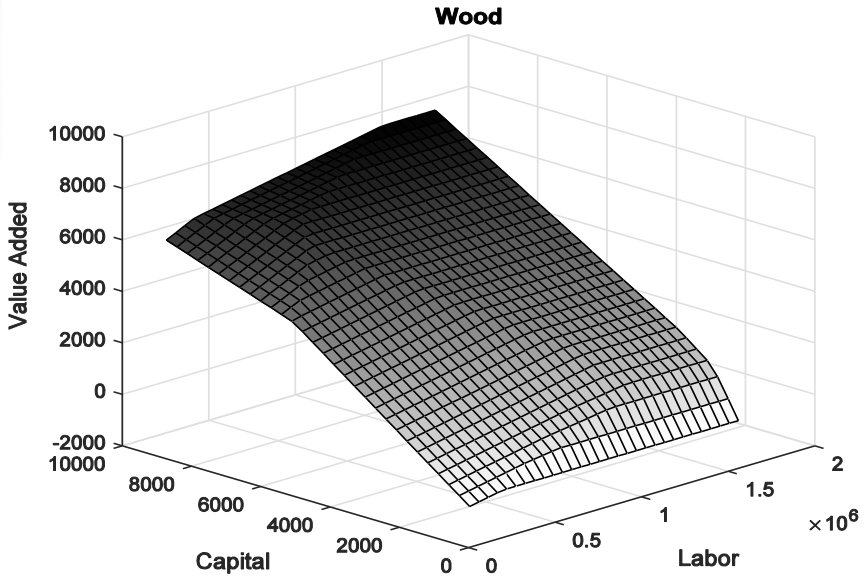


Table 14. Coefficient of contextual variables

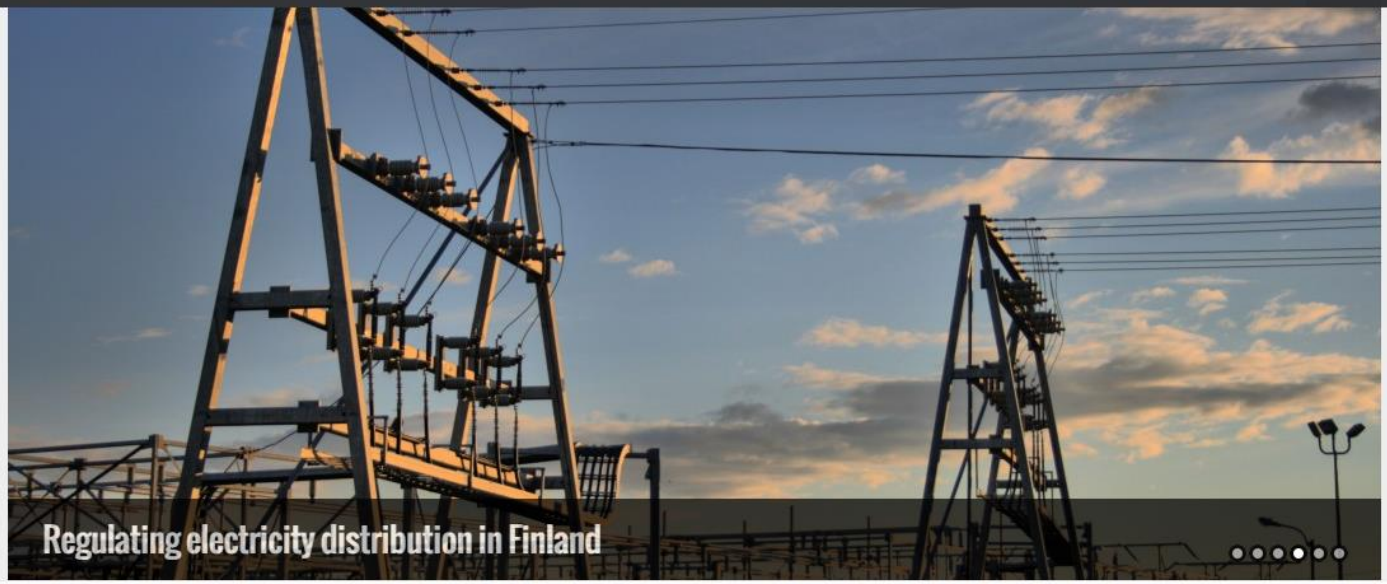
	Plastic (2520)		Wood (2010)	
	Dummy of exporting	Share of exporting in sales	Dummy of exporting	Share of exporting in sales
Point estimate	334.5	303.7	-763.0	4,114
95% lower bound	148.7	-334.3	-1944	2,568
95% upper bound	520.3	941.8	417.7	5,660
p-value	4.70×10^{-4}	0.349	0.203	5.64×10^{-7}

Insights

- Important to acknowledge survey or census application to test production function estimators
- Demonstrated a **model-selection framework** that assesses expected census error
- Application-Driven estimator selection framework **avoids extrapolation** of conclusions from simulated data.
- Many methods do similarly as well using real data

Andrew (Andy) L. Johnson, Ph.D.

Texas A&M University



News

Plenary Talk at European Workshop on Efficiency and Productivity Analysis (EWEPA) XIV

June 26th, 2015

The largest conference in the field of efficiency and productivity analysis is the European Workshop on Efficiency and Productivity Analysis (EWEPA) [...]

Informs Annual Conference - Nov 1-4 Philadelphia PA - DEA Cluster

June 25th, 2015

Seminars/presentation

- November 9th – Informs Annual Conference: A Multivariate Seminonparametric Bayesian Concave Regression Method to Estimate Stochastic Frontiers

  This presentation discusses a method that incorporates the latest advances in the Bayesian constrained regression literature offering an alternative to the current Least Squares-based and Kernel Regression-based Stochastic frontier constrained estimation methods, both in terms of runtime and of data capacity.

- October 4 and 5: College Industry Council on

Ongoing work

- Multi-variate Bayesian Convex Regression with Inefficiency

  This research builds in Hannah and Dunson's Multi-variate Bayesian Convex Regression to develop a method to estimate a shape constrained production functions and potential deviations from the function representing inefficiency.

- Shape Restricted Estimation of the Power Curve for a Wind Turbine

  The estimation of the power curve provides an application for methods to estimate production

Questions?

Johnson Laboratory Members