

Deep Learning for Econometrics

Greg Lewis (MSR + NBER) Matt Taddy (MSR + Chicago)

Jason Hartford (UBC) Kevin Leyton-Brown (UBC)

Kui Tang (Columbia) Dave Blei (Columbia)

Matt Goldman (MSFT) Justin Rao (MSR) Di Wang (MSFT)

James Zou (Stanford) Mengting Wan (UCSD) Richard Li (UW)

What do economists do?

320

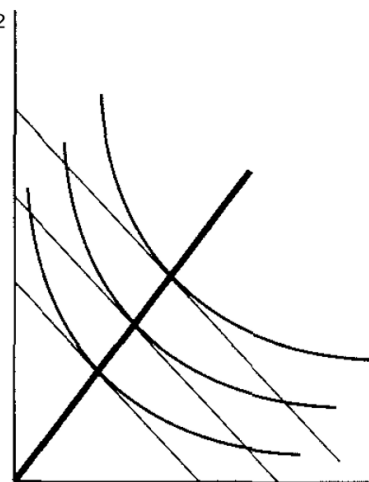
JUNE 1980

THE AMERICAN ECONOMIC REVIEW
TABLE 2—TOTAL EXPENDITURE AND OWN-PRICE ELASTICITIES

	Levels Model		First-Differences Model	
	Unconstrained e_i	Homogeneous e_{ii}	Unconstrained e_i	Homogeneous e_{ii}
Food	0.21	0.07	0.04	0.17
Clothing	2.00	-0.92	2.83	2.92
Housing	0.30	-0.31	0.04	-0.94
Fuel	1.67	-0.28	1.00	-0.31
Drink and Tobacco	1.22	-0.60	1.37	0.00
Transport and Communication	1.23	-1.21	1.14	-1.23
Other goods	1.21	-0.72	2.03	-0.52
Other services	1.40	-0.93	1.03	-0.78

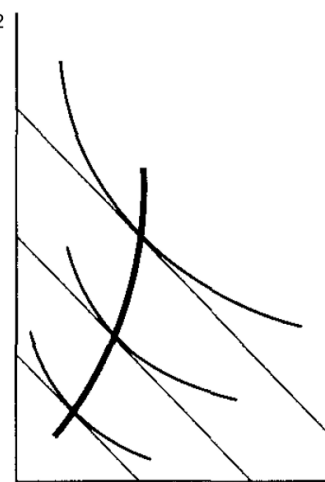
the D.W. statistic shows a sharp discussion of aggregation. A above, is that it assume that k , the distribution of household average structure budget. Finally, the separability

GOOD 2



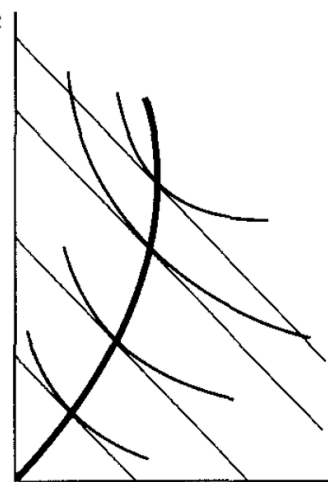
A

GOOD 2



B

GOOD 2



C

Income expansion paths. Panel A depicts unit elastic demands, in panel B good 2 is a luxury good, and in panel C, good 1 is an inferior good.

What do they need to do today?

The image is a composite of two screenshots. The background is a Google search results page for the query "toddler shoes". The search bar at the top shows the query and a magnifying glass icon. Below the search bar, there are tabs for "Web", "Images", "Videos", "Maps", "News", and "Explore". The "Web" tab is selected. The results show "91,800,000 RESULTS" and "Any time" filter. The first result is "Shop DSW Kids Shoes | dsw.com" with the URL "www.dsw.com/kids" and the text "DSW, Inc. The Latest Kids Styles @ Participating DSW Stores Today! ... and best brands for infants, toddlers, and ... ing Shoes, Sneakers, Dress Shoes | DSW". Below this, there are two promotional banners: "Sign Up for DSW® Rewards" (Earn a \$10 Certificate with Your First Purchase. Free to Enroll!) and "Find a Store Near You" (More Than 480 Locations Available. Shop at a DSW® Near You Today!).

Overlaid on the bottom left of the Google search results is a screenshot of the Microsoft Azure portal. The window title is "Microsoft Azure New HDInsight Cluster". The main content area is titled "Cluster Type configuration" and includes a "Report a bug" button. Below the title, there is a link "Learn about HDInsight and cluster versions. Learn more". The configuration options include "Cluster Type" (set to "Spark"), "Cluster Tier" (set to "STANDARD"), "Operating System" (set to "Linux"), and "Version" (set to "Spark 2.0.0 (HDI 3.5)"). The "STANDARD" tier is selected, showing features like "Administration" (Manage, monitor, connect), "Scalability" (On-demand node scaling), and "99.9% Uptime SLA". The "PREMIUM (PREVIEW)" tier is also visible, showing "Administration" (Manage, monitor, connect) and "Scalability" (On-demand node scaling).

What can we (AI) do to help?

The **dimension** of the economist's problem space has exploded

We can develop ML to navigate this space: **stay safe and automate**

We can also build new econometrics via deep structure

Example: Demand System

Suppose that you have transactions ' t ' on products ' j '.

Write the quantity bought ' q ' as

$$q_{tj} = \alpha_j(\mathbf{d}_t) + \gamma_j \log p_{tj} + e_{tj}$$

a function of utility we can ($\alpha_j(\mathbf{d}_t)$) and can't (e_{tj}) see, plus price p_{tj} .

You need to have a model like this to target customers or set prices.

But it's a system!

For example: There many different products

Demand for j depends on **substitutes** and **complements**

Or: where does price come from?

$$\log p_{tj} = \varphi_t(\mathbf{c}_j) + \psi_j q_{tj}^* + v_{tj}$$

and the *demand system* is in equilibrium when $q_{tj}^* = q_{tj}$

This equilibrium introduces `price endogeneity': $\mathbb{E}[p_{tj} e_{tj}] \neq 0$

Sometimes it's just regression

If we treat ε_{tj} as independent this is a prediction problem
e.g., model store transactions with covariates \mathbf{x}_{tj} as

$$\mathbb{E} \log q_{tj} = \mathbf{x}'_{tj} \boldsymbol{\beta} + \log p_{tj} \mathbf{x}'_{tj} \boldsymbol{\gamma}$$

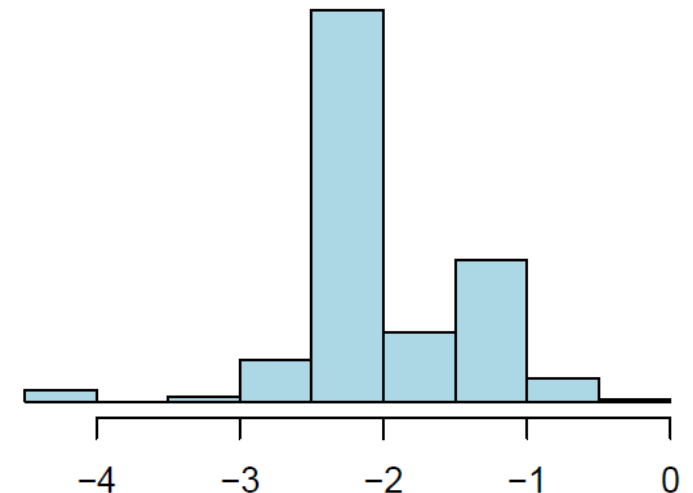
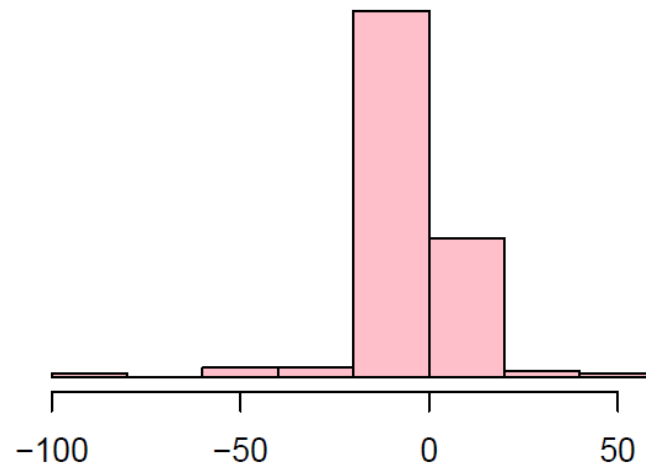
Elasticities:

one shared: $x_{tj} = 1$

brand-specific: $x_{tjk} = \mathbb{1}_{[k=j]}$

\mathbf{x}_{tj} = featurized description

$$\frac{dq}{dp} \frac{p}{q} = -0.23$$



Moving inside a demand system (AID)

It's *almost* ideal:

$$\mathbf{s}_t = \boldsymbol{\alpha} + \boldsymbol{\Gamma} \log(\mathbf{p}_t) + \boldsymbol{\beta} \log \frac{e_t}{\phi_t} + \boldsymbol{\varepsilon}_t$$

s_{tj} is the **budget share** for product j in basket t and e_t is the budget
($e_t = \sum_j \$_{tj}$ and $s_{tj} = \$_{tj}/e_t$)

ϕ_t is the **translog price index** $\sum_j \log p_{tj} [\alpha_j + \sum_k \gamma_{jk}^* \log p_{tk}]$

This is meaningful after aggregation, and **we can actually estimate it**

Factorizing Γ

The price terms are key to finding complements and substitutes

$$\mathbb{E} s_{tj} = \alpha_j + \sum_k \gamma_{jk} \log p_{tj} + \beta_j \frac{e_t}{\phi_t}$$

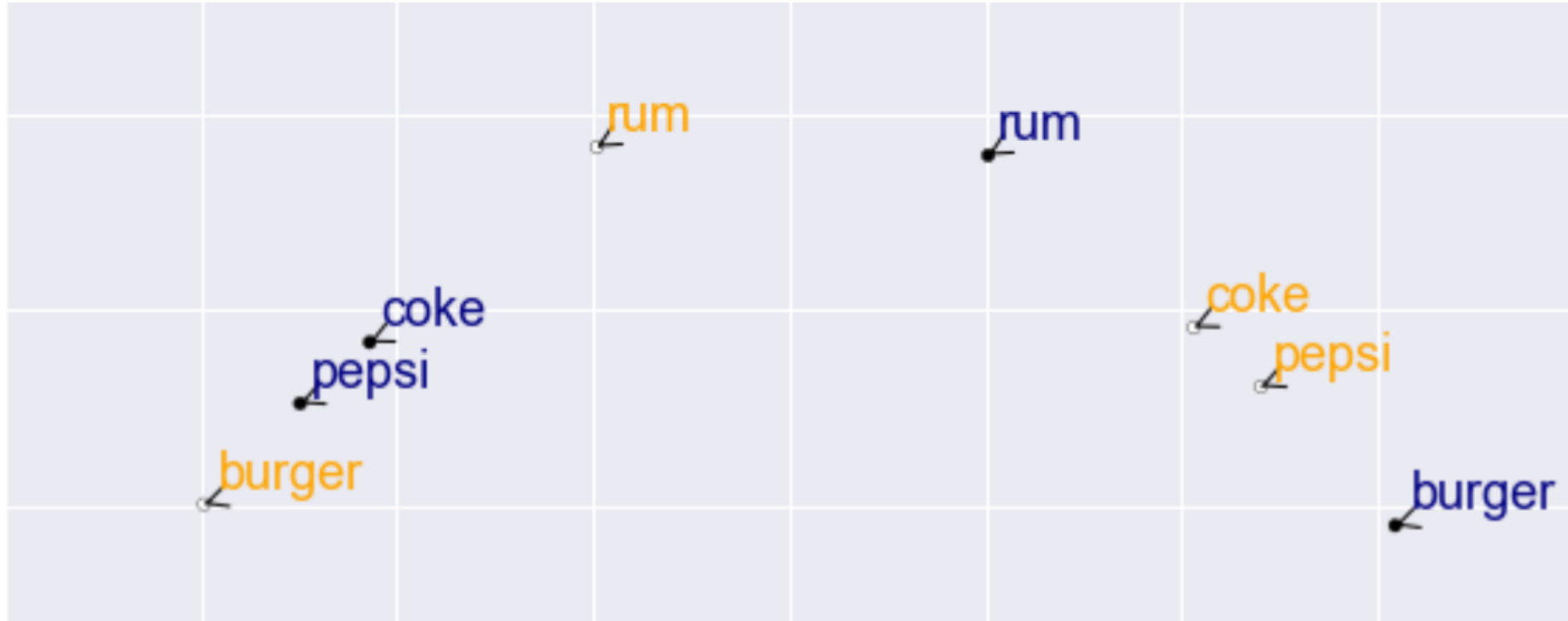
Γ is $J \times J$, so we need to reduce dimension if J is going to go big

One option: square matrix factorizations from word/prod embedding

$$\mathbf{\Gamma} = \mathbf{UV}' + \mathbf{VU}' + \mathbf{D} \text{ where } \mathbf{u}_j, \mathbf{v}_j \text{ are } S\text{-vectors and } \mathbf{D} \text{ is } J\text{-diagonal}$$

(AIDS implies restrictions: $\gamma_{jk} = \gamma_{kj}$, $\sum_j \gamma_{jk} = \sum_j \gamma_{kj} = \sum_j \beta_j = 0$)

Product Embeddings



substitutes (synonyms) are close in the same vector space
complements (topical words) are close across vector spaces

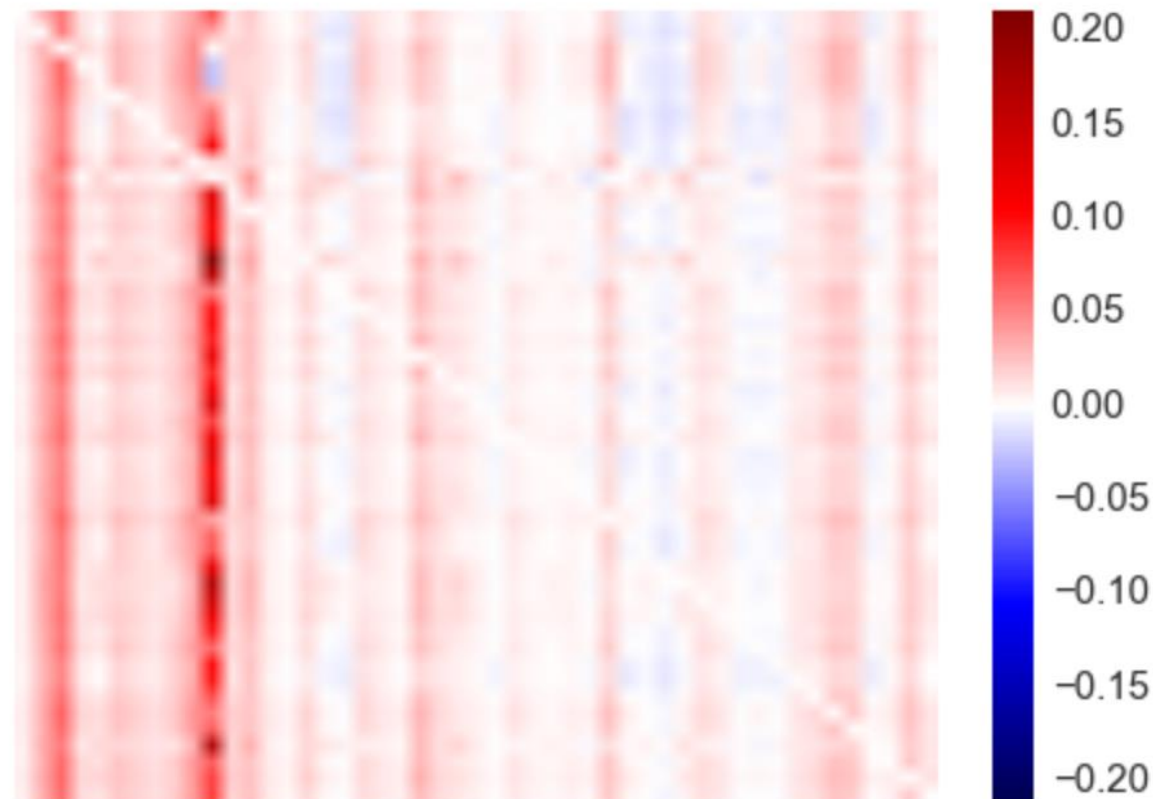
Beer

We fit on store-week totals.

Translate the γ_{jk} values into [compensated] elasticities as

$$\frac{\gamma_{jk}}{\bar{s}_j} - \bar{s}_k - \mathbb{1}_{[k=j]}$$

Elasticity matrix (omitting diagonal)



But wait... it's still a system

$$s_t = \alpha + \Gamma \log(\mathbf{p}_t) + \beta \log \frac{e_t}{\phi_t} + \mathbf{e}_t$$

Recall: where does price come from?

$$\log p_{tj} = \varphi_t(\mathbf{c}_j) + \psi_j q_{tj}^* + v_{tj}$$

and the *demand system* is in equilibrium when $q_{tj}^* = q_{tj}$

This equilibrium introduces 'price endogeneity': $\mathbb{E}[p_{tj} e_{tj}] \neq 0$

Endogenous Errors

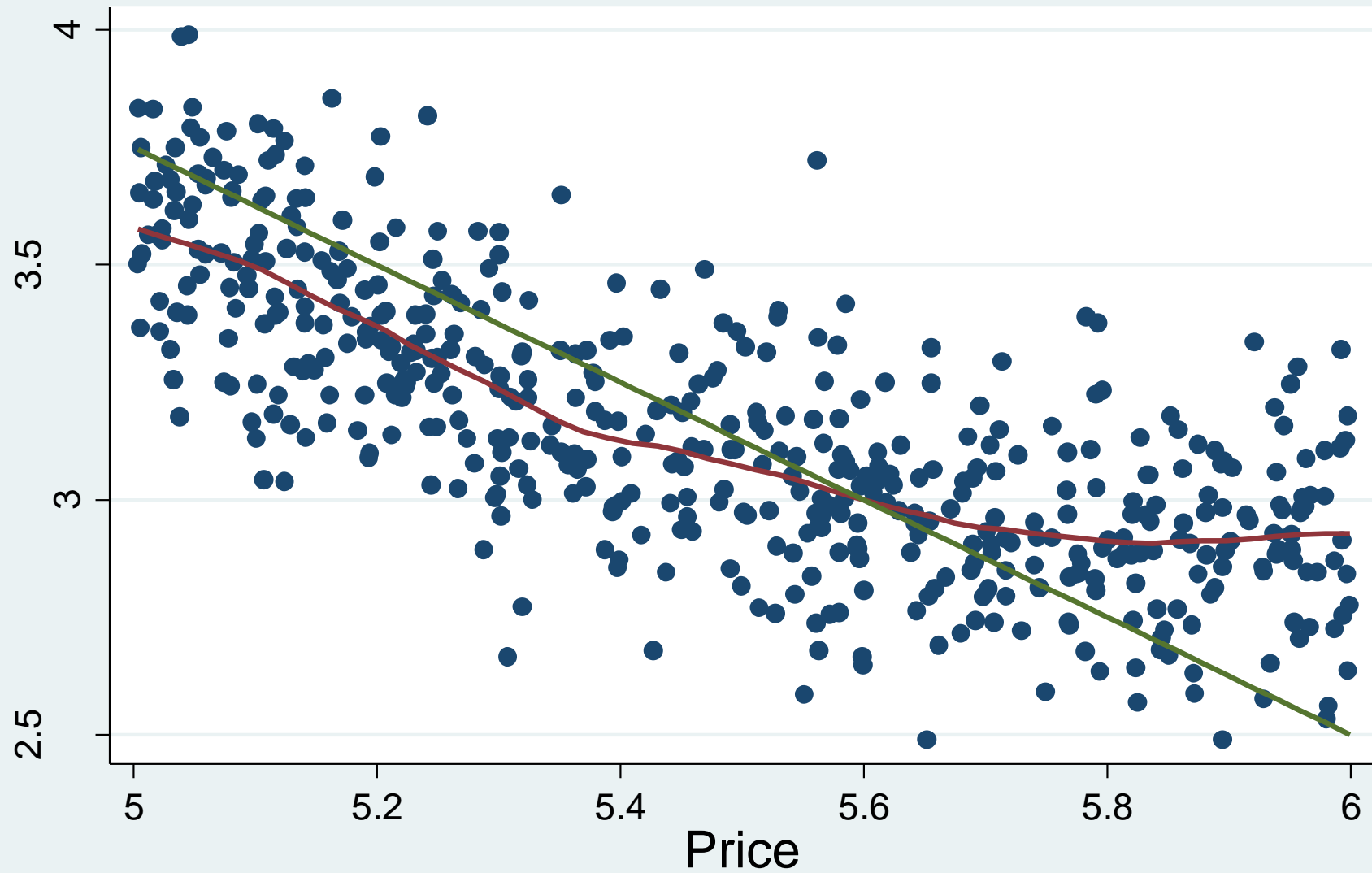
$$y = g(p, \mathbf{x}) + e \text{ and } \mathbb{E}[p e] \neq 0$$

If you estimate this using naïve ML, you'll get

$$E[y|p, \mathbf{x}] = E_{e|p}[g(p, \mathbf{x}) + e] = g(p, \mathbf{x}) + E[e|t, \mathbf{x}]$$

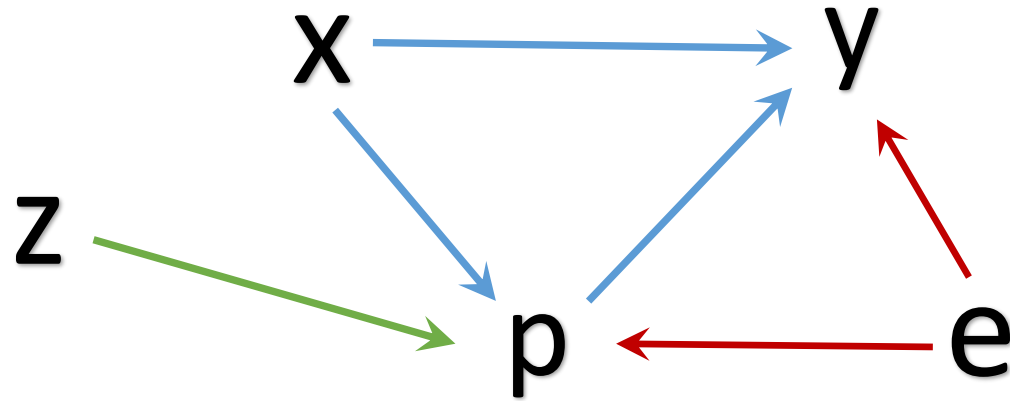
This works for **prediction**. It doesn't work for **counterfactual** inference:

What happens if I change p independent of e ?



— Prediction — Counterfactual Prediction

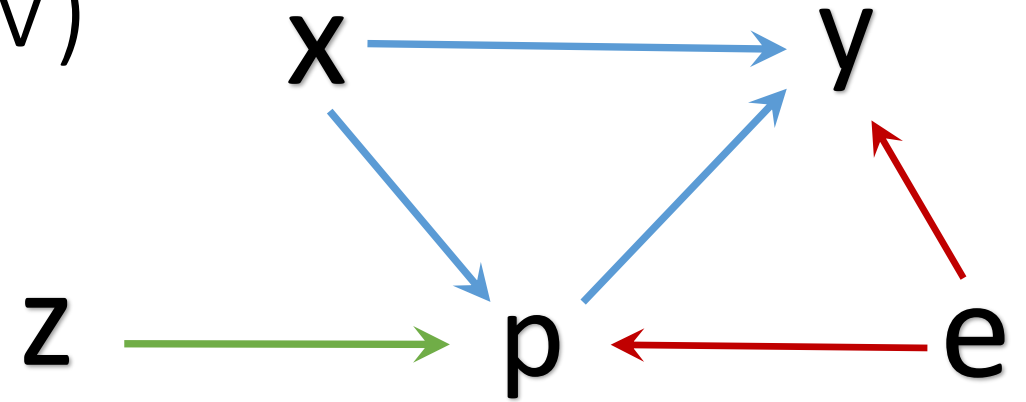
Instrumental Variables (IV)



In IV we have a special $z \perp e$ that influences policy p but not response y .

- Supplier costs that move price independent of demand (e.g., fish, oil)
- Any source of treatment randomization (intent to treat, AB tests, lottery)

Instrumental Variables (IV)



The *exclusion structure* implies

$$E[y|x, z] = E[g(p, x) + e|x, z] = \int g(p, x) dP(p|x, z)$$

So to solve for *structural* $g(p, x)$ we have a new learning problem

$$\min_{g \in G} \sum \left(y_i - \int g(p, x_i) dP(p|x_i, z_i) \right)^2$$

$$\min_{g \in G} \sum \left(y_i - \int g(p, x_i) dP(p|x_i, z_i) \right)^2$$

You might have seen 2SLS:

$$p = \beta z + v \text{ and } g(p) = \tau p \text{ so that } \int g(p) dP(p|z) = \tau \hat{p} = \tau \hat{\beta} z$$

So you first regress p on z then regress y on \hat{p} to recover $\hat{\tau}$.

This requires strict assumptions and homogeneous treatment effects.

$$\min_{g \in G} \sum \left(y_i - \int g(p, x_i) dP(p|x_i, z_i) \right)^2$$

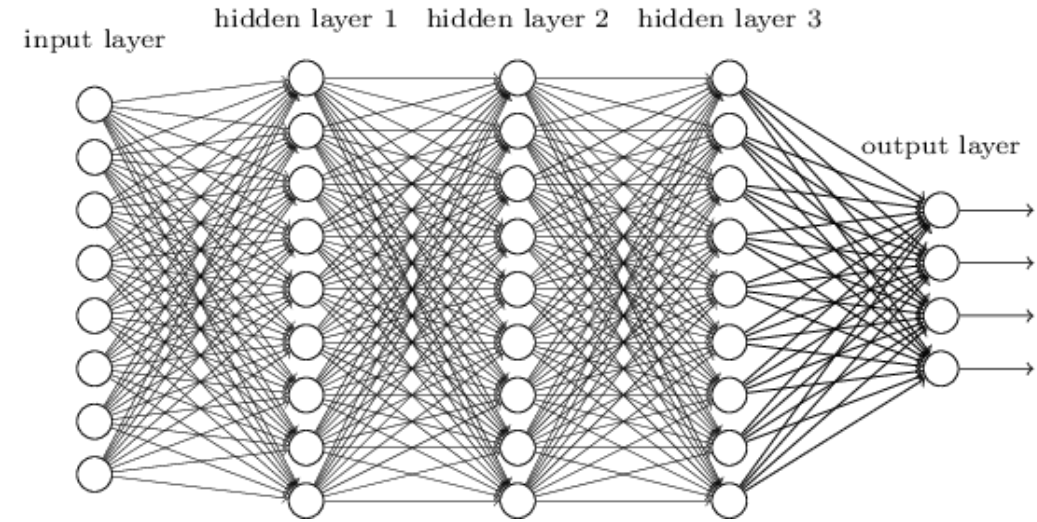
We can target this integral loss function directly with flexible g and P .

Brute force version

- Fit conditional distributions $\hat{P}(p|x_i, z_i)$.
- Generate $\{\hat{p}_{ib}\}_{b=1}^B \sim \hat{P}(p|x_i, z_i)$ for each i .
- Train \hat{g} to minimize $[y_i - B^{-1} \sum_b g(\hat{p}_{ib}, x_i)]^2$.

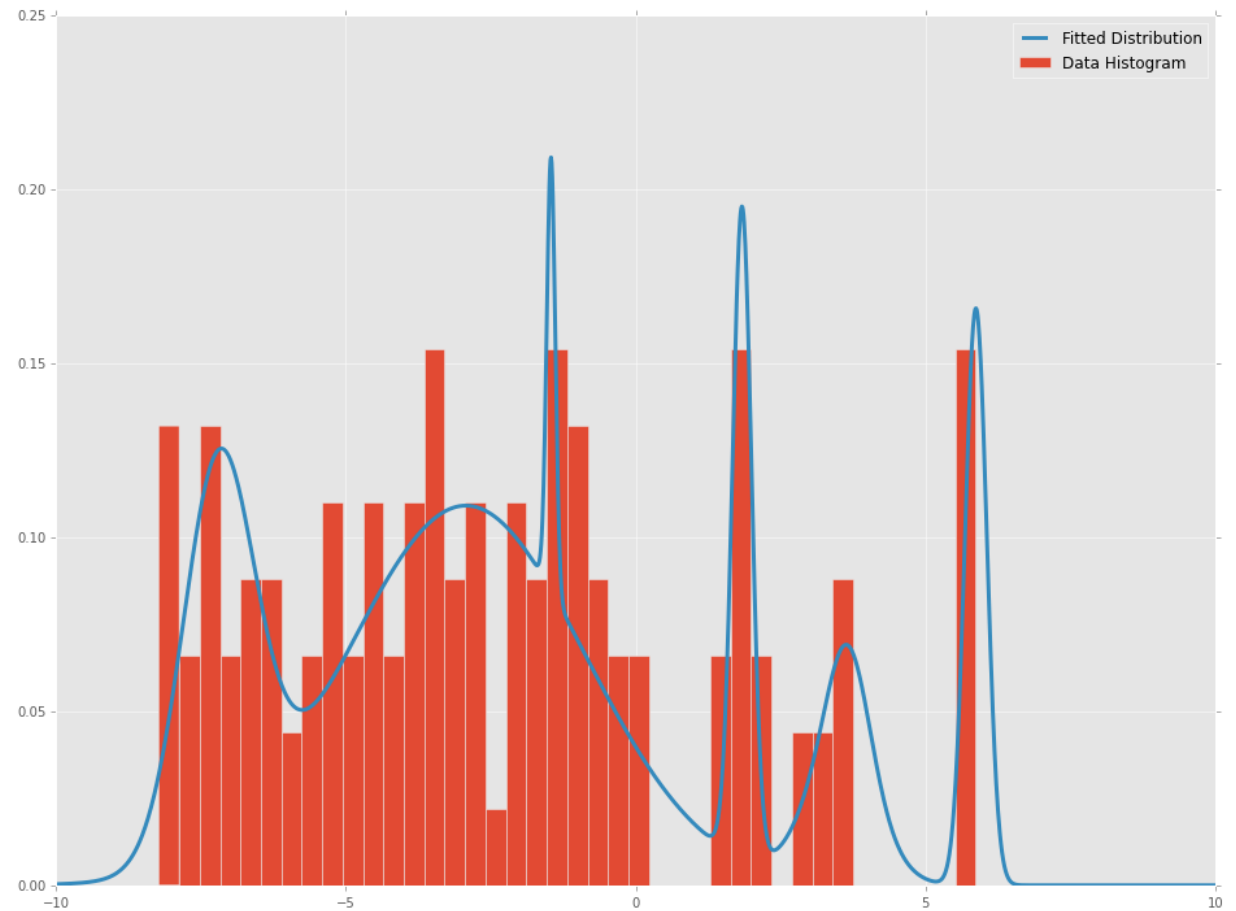
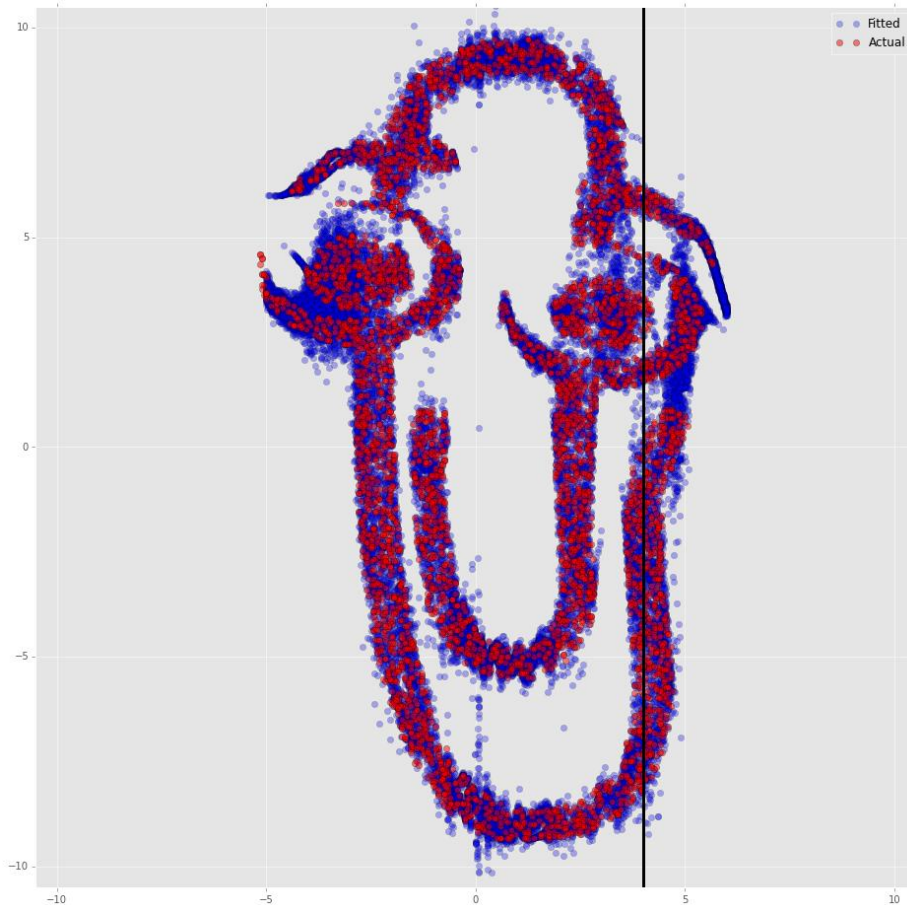
Turns IV into two ML tasks: we can use DNNs for both \hat{P} and \hat{g} .

Learning to love Deep Nets



First Stage is out-of-the-box ML: learn $P(p|x_i, z_i)$

e.g., DNN fits distribution to maximize likelihood for a mixture of Gaussians.



The second stage involves an integral loss function

Brute force just samples from $\hat{P}(p|x_i, z_i)$ to evaluate

$$\min_{\theta \in \Theta} \frac{1}{N} \sum_i \left(y_i - \frac{1}{B} \sum_b g(\hat{p}_{ib}, x_i; \theta) \right)^2, \quad \hat{p}_{ib} \sim \hat{P}(p|x_i, z_i)$$

Instead, *Stochastic Gradient Descent*: optimize via *unbiased* gradient estimates based upon mini-batch sample of the full dataset.

We can do SGD by pairing each observation with *two* treatment draws

$$\nabla g(\theta) \approx (y_i - g(\hat{p}_{i1}, x_i; \theta)) g'(\hat{p}_{i2}, x_i; \theta)$$

Linear Demand, Heterogeneous Effects

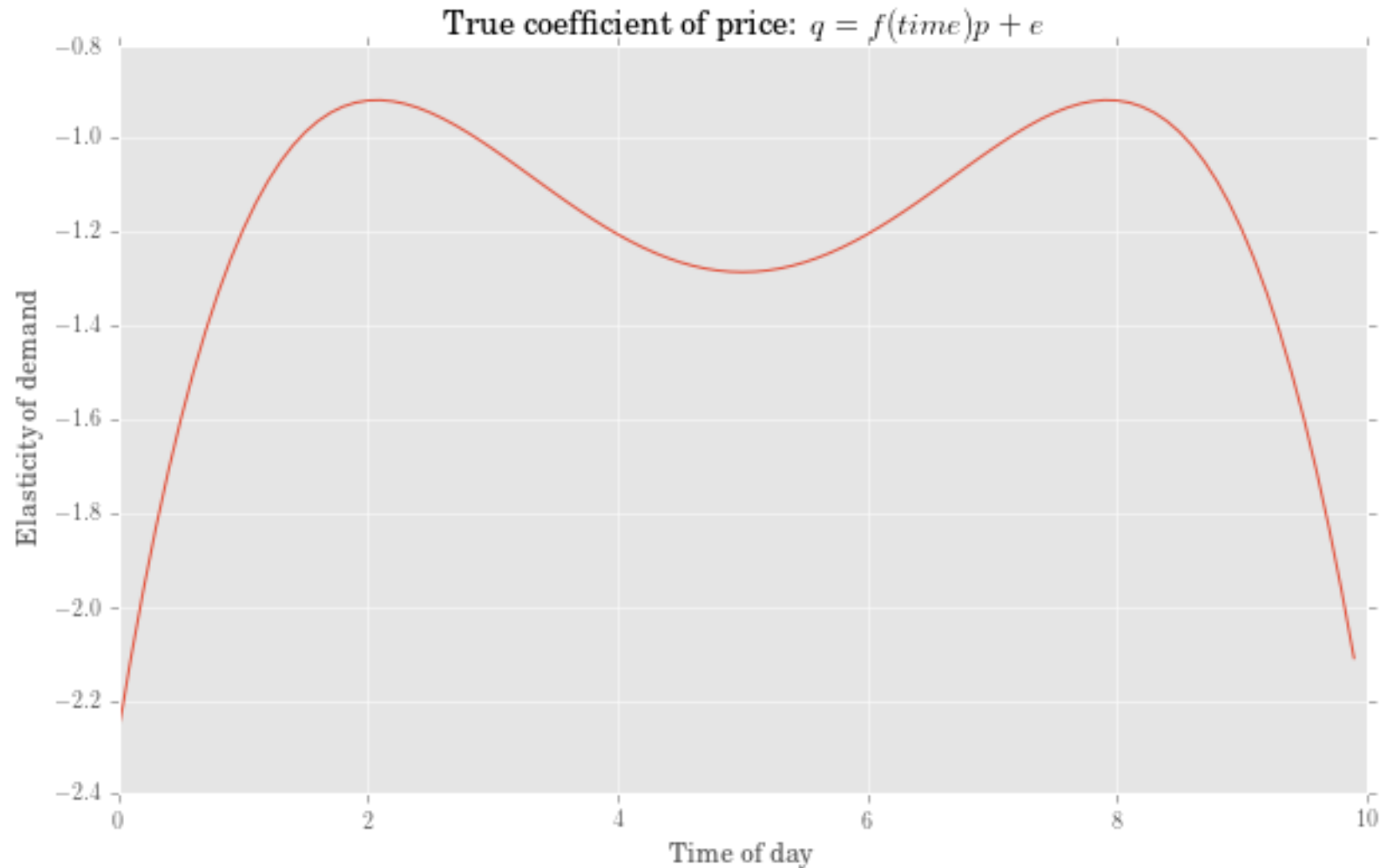
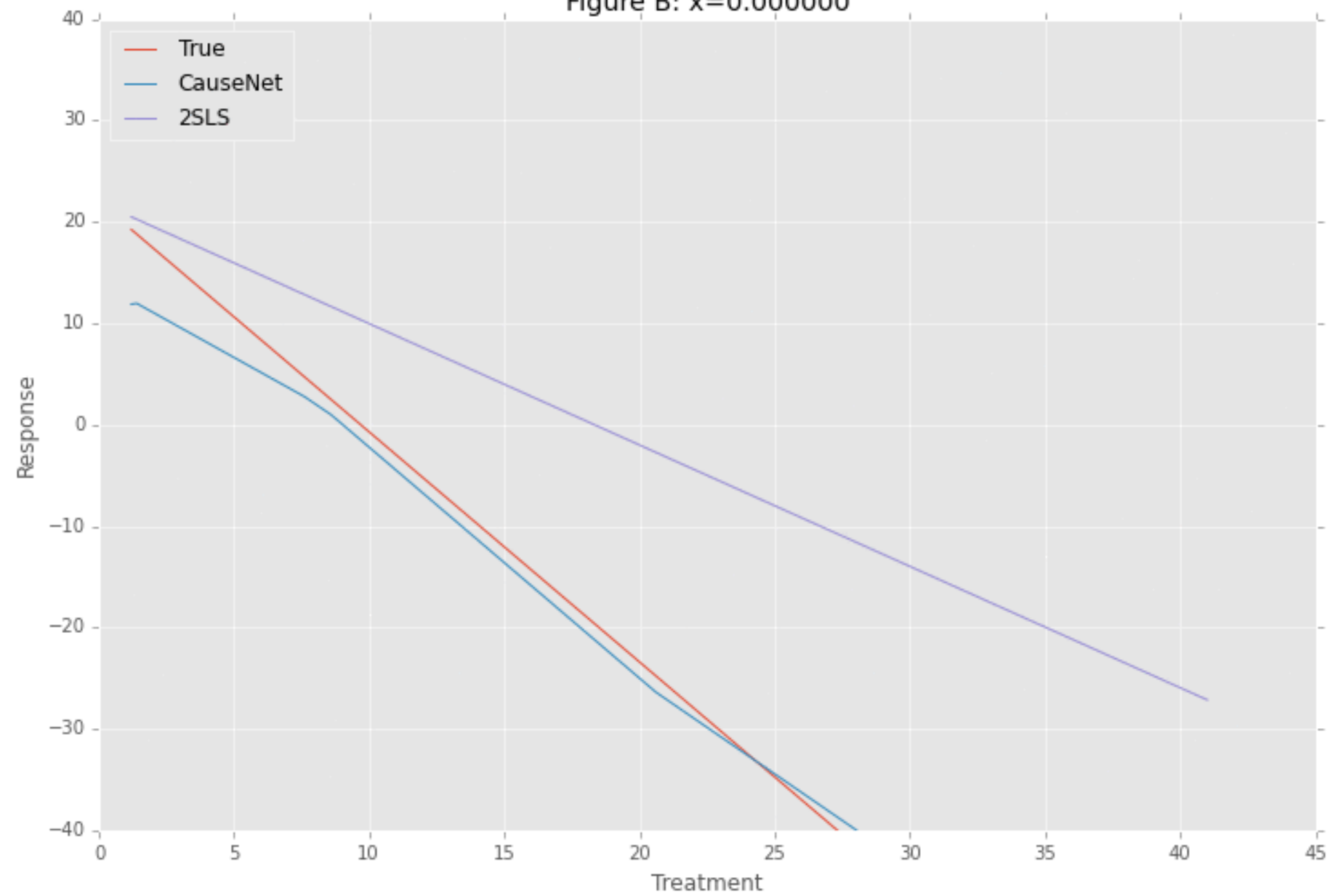


Figure B: $x=0.000000$



Ads Application

Taken from Goldman and Rao (2014)

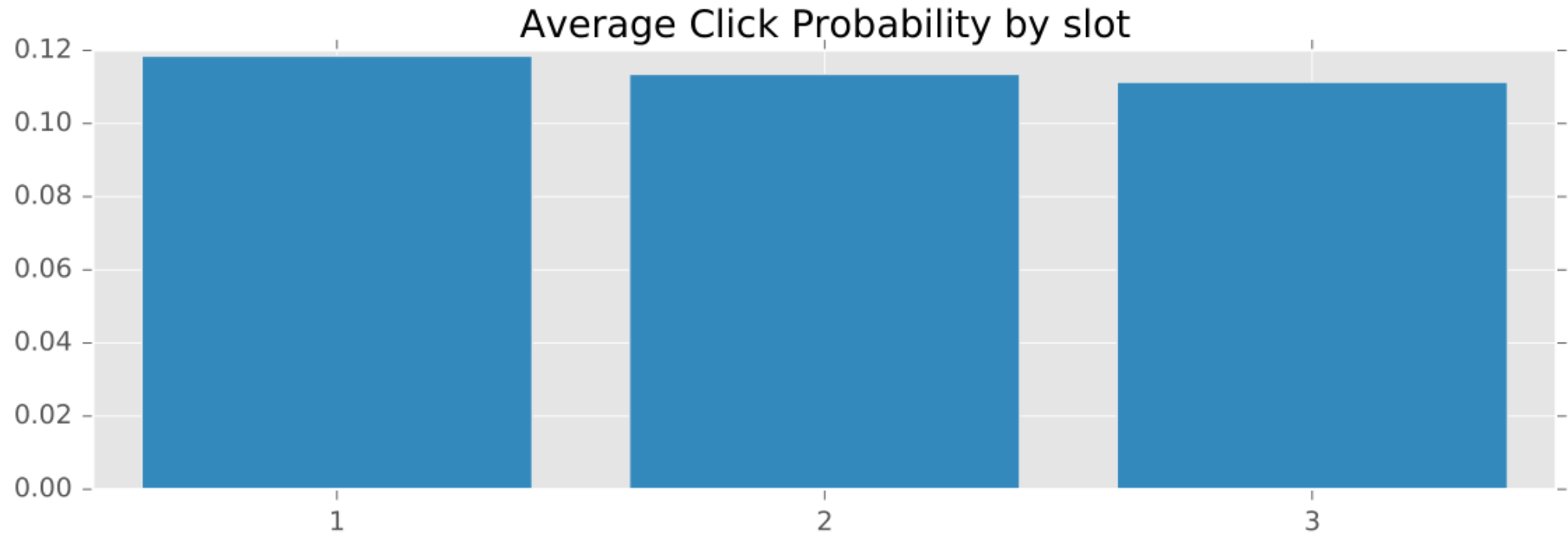
We have 74 mil click-rates over 4 hour increments for 10k search terms

Treatment: **ad position 1-3**

Instrument: **background AB testing**

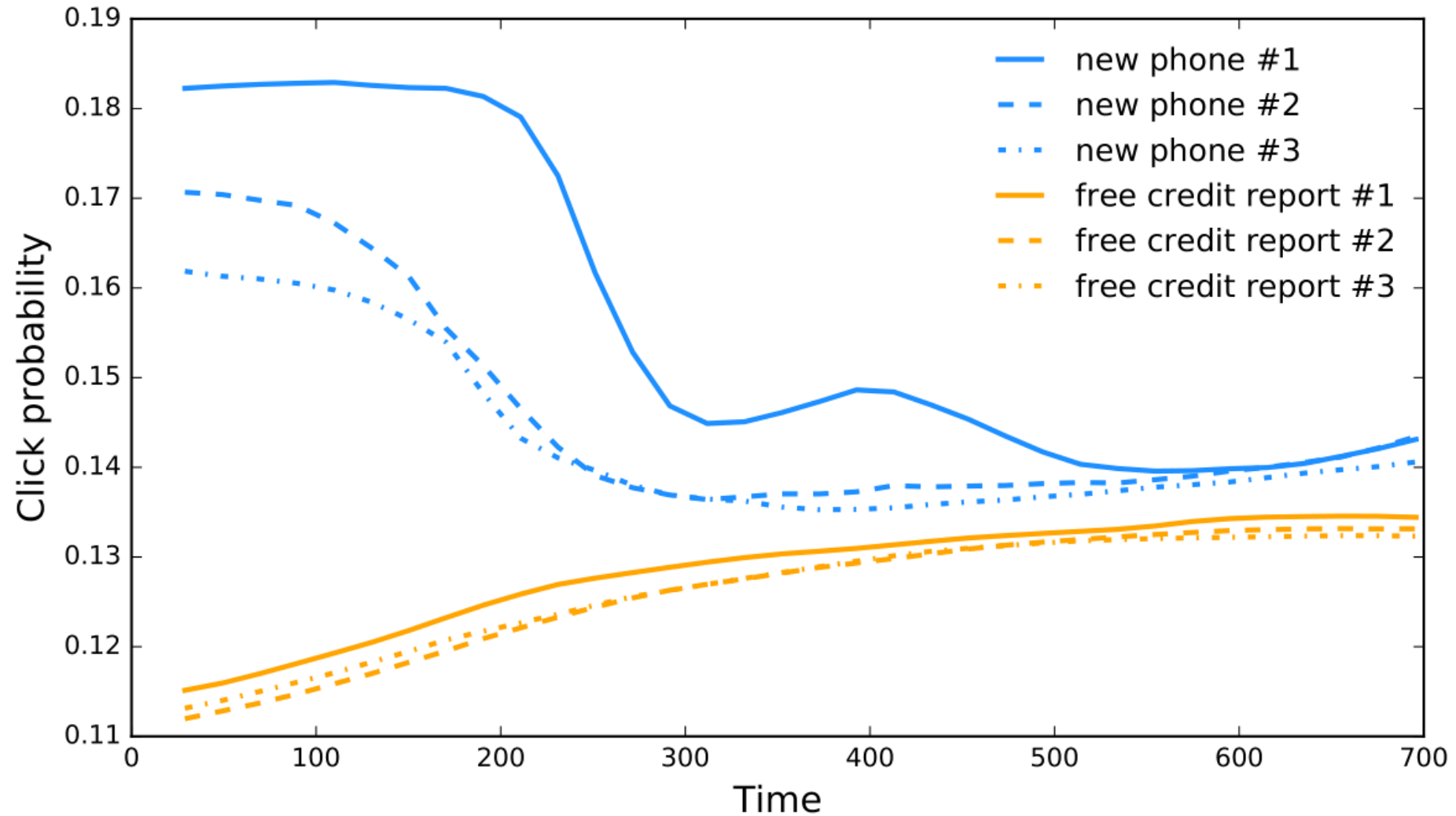
Covariates: **search text and time**

Average Treatment Effects



These compare to observed click probabilities of 0.33, 0.1, and 0.05.

Heterogeneous Treatment Effects



Economics and Artificial Intelligence

We have a track record pointing ML at questions of science + causation.

We're going to replicate this success at scale on unstructured data

We use economic theory to build systems of tasks that can be addressed with Deep nets and other state-of-the-art ML.

This is the construction of systems for Artificial *Economic* Intelligence.