The Demand for Data in Healthcare AI

Qihong Ruan

May 7, 2024

Executive Summary

Objective and Significance: This research proposal aims to estimate the demand functions for data used in training and fine-tuning large language models (LLMs) in China and the US, with a special focus on healthcare applications. As LLMs become increasingly important in various domains, including healthcare, understanding the demand for data is crucial for assessing the competitive dynamics, economic implications, and potential societal impacts of the rapidly evolving AI industry.

In the healthcare sector, LLMs have shown promise in applications such as clinical decision support, patient engagement, and medical research Rajkomar et al. [2019]. Accurate estimation of the demand for healthcare-related data can help policymakers, healthcare providers, and AI firms make informed decisions regarding data collection, sharing, and usage policies, ultimately contributing to the development of more effective and equitable AI-driven healthcare solutions.

Data Requirements and Methodology: The study assumes a "data fairy" setting, where researchers have access to comprehensive data on data offerings, firm-level data choices, instrumental variables, and potential market sizes. This rich data set allows for the application of state-of-the-art demand estimation techniques, specifically the random coefficients logit demand model, which accommodates flexible substitution patterns and heterogeneous preferences across LLM developers.

The estimation procedure involves three key steps:

1. Constructing market-level moments by matching observed market shares to those predicted by the model.
2. Creating micro-moments based on firm-level data to relate firm characteristics to data choices and help identify the distribution of random coefficients.
3. Using instrumental variables, such as the BLP instruments, to address the endogeneity of data prices and market shares.

Identification Strategy and Economic Intuition: The BLP instrumental variables approach is central to the identification strategy in this research. By exploiting exogenous variation in the characteristics of competing data products, the BLP instruments help identify the demand parameters and address the endogeneity of prices.

To illustrate the economic intuition behind the BLP instruments, consider an example with three healthcare-related datasets:

- Dataset A: Price = \$500, Size = 10,000 patient records, Quality = High
- Dataset B: Price = \$750, Size = 20,000 patient records, Quality = Medium
- Dataset C: Price = \$600, Size = 15,000 patient records, Quality = Medium

The BLP instruments for Dataset A could be constructed as follows:

- Sum of sizes of Datasets B and C: 20,000 + 15,000 = 35,000 patient records
- Average quality of Datasets B and C: (Medium + Medium) / 2 = Medium

These instruments capture the exogenous variation in the characteristics of competing datasets (B and C) that should affect the markup and market share of Dataset A without being correlated with its unobserved quality. By using these instruments, researchers can disentangle the effects of price and quality on demand, enabling a more accurate estimation of demand elasticities and welfare effects in the context of healthcare data for LLMs.

Expected Outcomes and Contribution: The research will provide robust estimates of the demand primitives for data in the context of LLMs, with a special emphasis on healthcare applications. These estimates can be used to calculate own- and cross-price elasticities, marginal costs, and markups, as well as to perform counterfactual simulations assessing the welfare effects of alternative market structures or policy interventions.

This study contributes to the growing literature on the economics of AI and digital markets, as well as the literature on the application of AI in healthcare Rajkomar et al. [2019], Topol [2019]. The insights gained from this research can help policymakers, healthcare providers, and AI firms navigate the complex landscape of data demand and AI-driven healthcare solutions, ultimately contributing to the development of more effective, efficient, and equitable healthcare systems.

1 Econometric Modeling and Identification

We employ a random coefficients logit demand model Berry et al. [1995], Nevo [2001] to estimate the demand for healthcare-related data in the context of large language models (LLMs). The indirect utility of healthcare provider (or AI firm) i for data product j in market t is given by:

$$u_{ijt} = \alpha_i p_{jt} + x_{jt}\beta_i + \xi_{jt} + \epsilon_{ijt}, \tag{1}$$

where p_{jt} is the price of data product j in market t, x_{jt} is a vector of observed data characteristics (e.g., size, quality, domain-specificity), ξ_{jt} represents unobserved data quality, and ϵ_{ijt} is an idiosyncratic taste shock. The random coefficients α_i and β_i capture provider-specific preferences for price and other data characteristics, respectively. The market share of data product j in market t is given by:

$$s_{jt} = \int \frac{\exp(u_{ijt})}{\sum_{k=1}^{J_t} \exp(u_{ikt})} dF(\alpha_i, \beta_i), \tag{2}$$

where J_t is the total number of data products in market t, and $F(\alpha_i, \beta_i)$ is the joint distribution of random coefficients. To estimate the model parameters, we

rely on the following micro-moments:

$$E[Z_{jt} \cdot (\xi_{jt}(\theta))] = 0, \tag{3}$$

where Z_{jt} is a vector of instrumental variables, $\xi_{jt}(\theta)$ is the unobserved data quality as a function of the model parameters θ, and $E[\cdot]$ denotes the expectation operator. The instrumental variables Z_{jt} should be uncorrelated with the unobserved data quality ξ_{jt} but correlated with the endogenous variables (prices and market shares). We employ the BLP instrumental variables approach Berry et al. [1995] to address the endogeneity of prices and market shares. The BLP instruments are constructed using the characteristics of competing data products:

- Sums and averages of competitor data characteristics: $\sum_{k \neq j} x_{kt}$ and $\frac{1}{J_t - 1} \sum_{k \neq j} x_{kt}$
- Number of competing data products: $J_t - 1$

These instruments capture the exogenous variation in the competitive environment that affects the markups and market shares of each data product without being correlated with its unobserved quality. In addition to the BLP instruments, we can leverage cost shifters and markup shifters as additional instrumental variables. Cost shifters, such as data collection and processing costs, affect prices without directly influencing demand. Markup shifters, like market structure and competition intensity, impact prices and market shares but are assumed to be uncorrelated with unobserved data quality. To estimate the model parameters, we use the generalized method of moments (GMM) Hansen [1982]. The GMM estimator minimizes the following objective function:

$$\min_{\theta} \xi(\theta)' Z W^{-1} Z' \xi(\theta), \tag{4}$$

where W is a positive definite weighting matrix, and $\xi(\theta)$ is the vector of unobserved data qualities as a function of the model parameters. The data required for estimation includes:

- Market-level data on healthcare-related data offerings, including prices, data characteristics, and market shares
- Provider-level data on data choices and characteristics
- Instrumental variables, such as BLP instruments, cost shifters, and markup shifters

By leveraging the BLP instrumental variables approach and the rich variation in the data, we can identify the demand parameters and obtain consistent estimates of price elasticities, substitution patterns, and welfare effects in the context of healthcare data for LLMs. This state-of-the-art econometric technique allows us to disentangle the effects of price and data quality on demand, providing valuable insights into the competitive dynamics and economic implications of the AI-driven healthcare industry.

2 Introduction

Estimating the demand functions for data in the context of large language models (LLMs) in China and the US is of significant importance for understanding the competitive dynamics and economic implications of the rapidly evolving AI industry. Data is a critical input for training and fine-tuning LLMs, and the demand for data is expected to grow as LLMs become more widely adopted Acemoglu [2021], Goldfarb and Tucker [2019], Agrawal et al. [2018].

The "data fairy" setting, which assumes access to any required data, allows us to leverage state-of-the-art demand estimation techniques and exploit rich variation in data characteristics, LLM performance, and market conditions. This ideal data scenario enables us to address the key challenges of demand estimation, such as endogeneity of data prices and the high dimensionality of data features Berry and Haile [2014], Berry et al. [1995], and to obtain precise and robust estimates of the demand primitives.

3 Data Requirements

To estimate the demand functions for data, we require the following:

- Market-level data on data offerings, including prices, data characteristics (e.g., size, quality, domain, labeling), and market shares for each data product in each market (defined as a combination of geography, time, and LLM application).
- Firm-level data on LLM developers' data choices, usage patterns, and firm characteristics (e.g., size, industry, technological capabilities) for a representative sample of firms in each market.
- Instrumental variables that shift data costs or markups without directly affecting demand, such as cost shifters (e.g., data collection and processing costs), markup shifters (e.g., market structure, competition), and BLP instruments Berry et al. [1995].
- Data on potential market sizes and outside options to accurately measure market shares and capture substitution to non-data alternatives (e.g., human-generated content or rule-based systems).

4 Demand Model

We specify a random coefficients logit demand model Berry et al. [1995], Nevo [2001] for data, which allows for flexible substitution patterns and heterogeneous preferences across LLM developers. The indirect utility of firm i for data product j in market t is given by:

$$u_{ijt} = \alpha_i p_{jt} + x_{jt}\beta_i + \xi_{jt} + \epsilon_{ijt} \tag{5}$$

where p_{jt} is the price, x_{jt} is a vector of observed data characteristics, ξ_{jt} represents unobserved (to the econometrician) data quality, and ϵ_{ijt} is an idiosyncratic taste shock. The random coefficients α_i and β_i capture firm-specific preferences for price and other data characteristics, respectively.

5 Estimation and Identification Strategy

We estimate the demand model using the following steps:

1. Construct market-level moments by matching observed market shares to those predicted by the model, as in Berry [1994]. This involves inverting the market share equations to obtain the mean utilities δ_{jt} that rationalize the observed shares.

2. Construct micro-moments based on firm-level data, as in Petrin [2002]. These moments relate firm characteristics to data choices and help identify the distribution of random coefficients.

3. Estimate the model parameters by GMM, using instrumental variables to address the endogeneity of data prices and market shares Berry et al. [1995]. Valid instruments include cost shifters, markup shifters, and BLP instruments (characteristics of competing data products).

4. Conduct post-estimation analyses, such as calculating own- and cross-price elasticities, marginal costs, and markups Nevo [2001], and perform counterfactual simulations to assess the welfare effects of alternative market structures or policy interventions.

To ensure identification, we leverage the BLP instrumental variables approach, which exploits exogenous variation in the characteristics of competing data products to address the endogeneity of prices.

5.1 BLP Instrumental Variables Example

Let's consider a simple example with three data products (A, B, and C) in a given market. Suppose we observe the following characteristics:

- Data Product A: Price = \$100, Size = 1TB, Quality = High
- Data Product B: Price = \$150, Size = 2TB, Quality = Medium
- Data Product C: Price = \$120, Size = 1.5TB, Quality = Medium

In this case, we could construct the following BLP instruments for Data Product A:

- Sum of sizes of Data Products B and C: 2TB + 1.5TB = 3.5TB
- Average quality of Data Products B and C: (Medium + Medium) / 2 = Medium

These instruments capture exogenous variation in the characteristics of competing data products (B and C) that should affect the markup and market share of Data Product A but are assumed to be uncorrelated with the unobserved quality of Data Product A (ξ_{At}).

With these BLP instruments, we can estimate the demand parameters using GMM, exploiting the moment conditions $E[\xi_{jt}|z_{jt}] = 0$, where z_{jt} represents the set of BLP instruments and other exogenous variables for data product j in market t. This approach allows us to identify the price coefficient and the distribution of random coefficients, capturing heterogeneous preferences across LLM developers.

6 Conclusion

Estimating the demand for data in the context of large language models in China and the US is a challenging but important task, with implications for our understanding of the AI industry and its societal impacts. The "data fairy" setting, combined with state-of-the-art demand estimation techniques and the BLP instrumental variables approach, allows us to overcome the key challenges and obtain robust estimates of the demand primitives.

This research contributes to the growing literature on the economics of AI and digital markets Agrawal et al. [2018], Goldfarb and Tucker [2019] and provides valuable insights for policymakers, firms, and researchers navigating this rapidly evolving landscape. By quantifying the demand elasticities, substitution patterns, and welfare effects associated with data in the context of LLMs, we can inform policy decisions, guide firm strategies, and shed light on the potential societal impacts of AI technologies.

References

Daron Acemoglu. Harms of ai. *National Bureau of Economic Research*, 2021.

Ajay Agrawal, Joshua Gans, and Avi Goldfarb. Prediction machines: the simple economics of artificial intelligence. 2018.

Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica*, pages 841–890, 1995.

Steven T Berry. Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, pages 242–262, 1994.

Steven T Berry and Philip A Haile. Identification in differentiated products markets using market level data. *Econometrica*, 82(5):1749–1797, 2014.

Avi Goldfarb and Catherine Tucker. Digital economics. *Journal of Economic Literature*, 57(1):3–43, 2019.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, pages 1029–1054, 1982.

Aviv Nevo. Measuring market power in the ready-to-eat cereal industry. *Econometrica*, 69(2):307–342, 2001.

Amil Petrin. Quantifying the benefits of new products: The case of the minivan. *Journal of political Economy*, 110(4):705–729, 2002.

Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019.

Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.