CERGE - EI

Centre for Economic Research and Graduate
Education - Economics Institute

CHARLES
UNIVERSITY

Estimating Genre-Specific Demand Elasticities in the Amazon Book Market: A Double/Debiased Machine Learning Approach

MASTER THESIS

Bc. Ladislav Ďurian

Supervisor of the Master Thesis: Paolo Zacchia, Ph.D.
Study Programme: Master in Economic Research

Prague 2026

I declare that I carried out this Master thesis on my own, and only with the cited sources, literature and other professional sources. During the preparation of this thesis, I used ChatGPT solely to improve the clarity and readability of some paragraphs. After using this tool, I reviewed and edited the content as necessary and take full responsibility for the content of the thesis.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In .... date ....

Author's signature

I would like to express my gratitude to my supervisor, Paolo Zacchia, Ph.D., for his expert guidance and honest feedback throughout the writing of this thesis. I also thank Michal Fabinger, Ph.D., for his advice on fine-tuning the neural network architecture, and doc. Nikolas Mittag, Ph.D., for his practical insights on the heterogeneity analysis. Finally, I would like to thank my family for their continued support and encouragement throughout my studies and the writing of this thesis.

Title: Estimating Genre-Specific Demand Elasticities in the Amazon Book Market: A Double/Debiased Machine Learning Approach

Author: Bc. Ladislav Ďurian

Institute: CERGE-EI

Supervisor: Paolo Zacchia, Ph.D.

Abstract: In this thesis, I examine how demand elasticities with respect to price, star ratings, and newspaper reviews vary across book genres in the Amazon book marketplace. Using daily panel data from Amazon domains in the United States, the United Kingdom, and Canada in 2018, I estimate a benchmark linear fixed-effects model and contrast it with a double/debiased machine learning specification that flexibly accounts for the nonlinear effects of control variables. I find substantial heterogeneity in consumer demand across genres and suggest how marketers can exploit this heterogeneity when designing genre-based segmentation strategies. Beyond its marketing implications, the thesis offers a methodological insight by showing that double/debiased machine learning yields genre-specific elasticity estimates that are robust to nonlinear confounding, unlike linear fixed-effects models.

Keywords: causal inference, causal machine learning, heterogeneous treatment effects, demand elasticity, Amazon book marketplace

Název práce: Odhad žánrově specifických elasticit poptávky na knižním trhu Amazonu: Metoda double/debiased machine learningu

Autor: Bc. Ladislav Ďurian

Pracoviště: CERGE-EI

Vedoucí diplomové práce: Paolo Zacchia, Ph.D.

Abstrakt: V této práci zkoumám, jak cena, hodnocení zákazníků a recenze v tisku ovlivňují elasticitu poptávky napříč knižními žánry na Amazonu. S využitím denních panelových dat z platforem Amazonu ve Spojených státech, Spojeném království a Kanadě z roku 2018 odhaduji benchmarkový lineární model s fixními efekty a porovnávám jej s metodou double/debiased machine learningu, která flexibilně modeluje nelineární efekty kontrolních proměnných. Výsledky ukazují značnou heterogenitu v poptávce spotřebitelů napříč žánry a navrhuji, jak mohou marketéři tuto heterogenitu využít při segmentaci trhu podle žánrů. Kromě přínosů pro marketing má práce i metodologický rozměr a ukazuje, že double/debiased machine learning dokáže odhadnout žánrově specifické elasticity bez skreslení i v přítomnosti nelineárních vztahů mezi proměnnými, na rozdíl od lineárních modelů s fixními efekty.

Klíčová slova: kauzální inference, kauzální strojové učení, heterogenní efekty treatmentu, elasticita poptávky, knižní trh Amazonu

Contents

Introduction

Marketing aims to understand consumer preferences to design and deliver appropriate goods and services. Marketers are interested in what products to offer, what prices to charge, how to promote the products, and how to deliver the products to the consumer. Consumer preferences are diverse, giving rise to market segments. As consumer preferences become more heterogeneous, treating the market in the aggregate becomes less efficient. Understanding the heterogeneity in consumer preferences is crucial for marketers, as it allows them to tailor differentiated product offerings for each segment (Allenby & Rossi, 1998).

In contrast to the emphasis on individual differences, economists are often more interested in average treatment effects (ATEs) and regard heterogeneity as a nuisance parameter problem, which they address using fixed effects (FE). While FE models allow heterogeneity in intercepts, in the marketing context, there is no reason to believe that distinctions in consumer behavior should be limited to intercepts, because differences in slope coefficients are critically important. Disregarding consumer diversity may inappropriately combine members from heterogeneous subpopulations, resulting in misleading treatment effect estimates. Econometricians have known for decades that random coefficient models accommodate slope heterogeneity (Swamy, 1970). Still, their practical estimation often relies on strong distributional assumptions that can be unrealistic in applied settings (Allenby & Rossi, 1998). The emergence of big data in economics has brought vastly larger panels with richer covariates, making it easier to estimate how slope coefficients vary across individuals. Among the most valuable sources of such data is the Amazon book marketplace, which accounted for around 45 percent of the U.S. physical book market in 2018 and serves as a uniquely powerful lens into consumer demand (Reimers & Waldfogel, 2021).

Using a high-frequency panel with daily data on book demand in 2018 from the Amazon domains in the United States, Canada, and the United Kingdom, I investigate heterogeneity in consumer demand in the Amazon book marketplace. I focus on the book market as book titles reflect the interests of specific population subgroups. As French novelist and Nobel Prize laureate François Mauriac once said, “Tell me what you read, and I’ll tell you who you are”, highlighting how reading choices can reveal aspects of identity. Inspired by his insight, I use book genres to represent different groups of readers on the market. Although my dataset allows me to explore heterogeneity at the level of editions, authors, or publishers, I focus on genres because they are a universally understood, broad category that points to distinct reader segments. Everyone understands what to expect from a horror versus a romance, while my results could be harder to grasp if my analysis centers on authors or titles that the general population is not familiar with. By analyzing genres, I can draw more generalizable insights into demand heterogeneity. Moreover, presenting results at the level of editions, authors, or publishers can be challenging because I have thousands of them in my dataset, compared to only about fifty genres. To my knowledge, there is no study that explores heterogeneity in consumer demand across book genres.

My thesis builds upon the paper by Reimers & Waldfogel (2021), which analyzes the aggregate effects of price, Amazon star ratings, and professional newspaper reviews on consumer demand in the Amazon book marketplace. I extend their work by showing how the demand elasticities with respect to price, star ratings, and newspaper reviews vary across book genres, following their approach of using Amazon sales rank as a proxy for consumer demand. Sales rank is a reasonable proxy for demand because the relationship between the natural logarithm of sales rank and the natural logarithm of demand volume is approximately linear (Chevalier & Goolsbee, 2003). This implies that the elasticities I estimate are sales rank elasticities that are proportional to demand elasticities up to a scalar transformation, allowing informative comparisons of relative demand responsiveness across genres.

To estimate genre-specific elasticities, I employ a conditional average treatment effects (CATEs) framework and extend the linear FE model of Reimers & Waldfogel (2021) by interacting the treatment variables with genre dummies. However, genre-specific elasticity estimates may be biased if variables interact nonlinearly. Notably, Reimers & Waldfogel (2021) acknowledge the presence of nonlinear relationships by including first- through third-order polynomials in the number of days until and since the publication of a book. However, this approach does not rule out the possibility of additional nonlinearities in other covariates or in their interactions. A tolerance for functional form errors shrinks when estimating CATEs. A model that yields unbiased ATEs can still give biased CATEs. This is because the treatment-genre interactions may pick up genre-specific nonlinearities that often average out in a pooled regression used to estimate ATE (Blackwell & Olson, 2022). To avoid misspecifying the functional form by assuming pure linear relationships, I employ the double/debiased machine learning (DML) methodology by Chernozhukov et al. (2018), which allows me to account for the nonlinear effects of confounding variables while retaining the interpretability of treatment effects as elasticities. I compare the DML estimates with those from the linear FE model to examine whether nonlinearities bias my genre-specific elasticity estimates. Estimating genre-specific elasticities without bias is crucial because these elasticities inform pricing and marketing decisions. Biased estimates can lead marketers to misdirect budgets toward ineffective campaigns within the targeted genres.

My results using either the linear FE or DML models indicate substantial heterogeneity across genres in how sales rank responds to prices, star ratings, and newspaper reviews. In particular, genre-specific CATEs differ markedly from ATEs, implying meaningful differences in demand responsiveness across genres. Furthermore, genre-specific CATE estimates for price, star ratings, and newspaper reviews from the linear FE model differ significantly for a substantial number of genres from the corresponding estimates in the DML model. This indicates that some genre-specific CATEs may have been distorted by unmodeled nonlinear relationships in the linear FE model that the DML approach can capture. Moreover, the DML specification typically delivers more precise genre-specific estimates than the linear FE model, consistent with efficiency gains from accounting for nonlinear confounding. Due to its ability to model nonlinear impacts of confounders and its higher precision in estimates, the DML model constitutes the main specification of this thesis.

These findings translate into actionable insights for booksellers and marketers designing genre-based segmentation strategies. For pricing, genres such as Philosophy, Psychology, and Cooking tend to respond strongly to price changes: even small price reductions are associated with noticeable improvements in sales rank, reflecting strong demand responsiveness to price changes. In contrast, genres such as Literary Criticism, Music, and Nature show only a small response to price changes, suggesting that modest price increases are unlikely to substantially reduce demand in these categories. For star ratings, genres such as Travel, Music, and Health & Fitness show large sales rank shifts with changes in star ratings. Marketers promoting books in these genres may benefit from emphasizing high star ratings, as demand is relatively responsive to changes in star ratings. In contrast, genres such as Action & Adventure, Fiction: Horror, and Philosophy do not significantly shift their sales rank with changes in star ratings. Highlighting high star ratings in marketing campaigns for these genres is therefore unlikely to produce meaningful changes in demand. For newspaper reviews, genres such as Comics & Graphic Novels, Philosophy, and Religion improve their sales rank noticeably after being reviewed by a newspaper. In these genres, highlighting that a book has been reviewed by a professional newspaper, for example, by placing review excerpts on the book cover, can significantly boost demand. In contrast, genres such as Music, Travel, and Cooking improve their sales rank only slightly after the appearance of a newspaper review. For these genres, emphasizing professional reviews on the book cover is unlikely to significantly impact demand.

The remainder of this thesis is organized as follows. Chapter 1 reviews the literature on consumer heterogeneity, heterogeneous treatment effects, and the use of machine learning in economics. Chapter 2 introduces the DML methodology, discusses identification and inference. Chapter 3 presents the empirical application, describing the empirical setting and the data. Chapter 4 outlines the empirical strategies, including the linear FE model and the DML approach. Chapter 5 reports the results, compares estimates across the two specifications, and visualizes the nonlinear confounding structure.

1 Literature Review

1.1 From Consumer Heterogeneity to Heterogeneous Treatment Effects

One of the biggest challenges in marketing is understanding the diversity of consumer preferences that exist in the market. As consumer preferences become increasingly heterogeneous, treating the market as a whole becomes less efficient. This heterogeneity gives rise to differentiated product offerings designed for each segment (Allenby & Rossi, 1998).

In contrast to the emphasis on individual differences in marketing, economics often focuses on average treatment effects (ATEs) and treats heterogeneity as a nuisance parameter, addressing it through fixed effects (FEs). While FE models allow for heterogeneity in intercepts, there is little reason to believe that differences in consumer behavior should be confined to intercept shifts, as differences in slope coefficients are frequently more consequential. Ignoring these differences can lead to the improper grouping of individuals from heterogeneous subpopulations, resulting in misleading treatment effect estimates. Although random coefficient models accommodate slope heterogeneity (Swamy, 1970), their practical implementation relies on strong distributional assumptions such as joint normality of individual-specific slope coefficients. Recognizing that consumers differ not only in baseline demand levels but also in their responsiveness to the price changes naturally motivates a shift from ATEs to the conditional average treatment effect (CATE) framework (Allenby & Rossi, 1998).

1.2 Average and Conditional Treatment Effects

A convenient way to formalize heterogeneity in treatment responses is through the potential outcomes framework (Rosenbaum & Rubin, 1983), which forms the foundation for the causal analysis. Let D be a dummy variable indicating treatment status, with $D = 1$ if an individual receives the treatment and $D = 0$ otherwise. Denote by $Y(1)$ the potential outcome for an individual if he receives the treatment, and by $Y(0)$ the potential outcome if he does not receive the treatment. The observed outcome is therefore $Y \equiv D \cdot Y(1) + (1 - D) \cdot Y(0)$. Let X be a vector of covariates capturing characteristics that may influence both treatment assignment D and outcomes Y.

The ATE summarizes the mean impact of the treatment in the population and is defined as:

$$ATE \equiv \mathbb{E}[Y(1) - Y(0)] \tag{1.1}$$

While useful as an aggregate measure, the ATE hides systematic differences in treatment responses across individuals or groups. To capture these differences, the CATE conditions on covariates:

$$CATE(x) \equiv \mathbb{E}[Y(1) - Y(0) \mid X = x] \tag{1.2}$$

Studying CATEs allows researchers to quantify how the treatment effect varies with observable characteristics (Abrevaya, Hsu, & Lieli, 2015). The emergence of big data in economics has brought much larger panels with richer covariates, facilitating CATE estimation as more observations with similar X are available, which improves the precision with which CATEs can be identified.

1.3 Machine Learning in Economics

1.3.1 Predictive Machine Learning in Economics

As computers became involved in economic transactions, the volume of data for analysis increased dramatically, leading to the emergence of big data in economics. Conventional statistical and econometric techniques, including linear regression, often work well with large datasets, but there are challenges unique to big data that may require different tools. First, the dataset may have more potential predictors than the number of observations, calling for variable selection. Second, large datasets may exhibit more complex relationships among variables that linear models may not be able to capture. Wooldridge (2020) emphasizes that the estimates in the linear model would be biased even if all relevant variables are included, but in the wrong functional form. Machine learning (ML) methods provide flexible tools for addressing both issues (Varian, 2014).

A commonly used technique for selecting relevant covariates when their number exceeds the number of observations, but only a subset actually affects the outcome, is the least absolute shrinkage and selection operator (LASSO). LASSO shrinks the coefficients of irrelevant predictors to zero, retaining only the relevant ones, which allows estimation to proceed (Tibshirani, 1996). By retaining only the most predictive variables, LASSO balances model complexity and generalizability. However, it performs best in sparse models, where only a few variables are good predictors, and it may have lower predictive power when many variables are equally relevant for prediction (Strittmatter, 2025).

Decision trees serve as an effective tool to capture nonlinear relationships in the data (Breiman, Friedman, Olshen, & Stone, 1984). A decision tree partitions the data into branches based on rules derived from input variables, sequentially splitting observations according to the variable and threshold that most effectively reduce prediction error at each step. Each branch terminates in a leaf, where the average outcome among the observations in that leaf represents the prediction. While decision trees can capture nonlinear patterns in a straightforward and interpretable way, they are sensitive to small changes in the sample, which can lead to instability in the tree structure. Random forests address this instability by averaging predictions across multiple decision trees, each trained on random subsamples of the data and covariates (Breiman, 2001). This aggregation improves predictive accuracy and reduces overfitting, but it comes at the cost of interpretability, because the resulting predictions are formed from potentially hundreds of trees rather than a single, transparent decision rule (Strittmatter, 2025).

Neural networks extend the ability to model nonlinear relationships far beyond what decision trees or random forests can typically capture. They consist of interconnected layers of neurons that transform the input data through learned weights and nonlinear activation functions, such as the sigmoid or hyperbolic tangent. As information passes through these layers, the network uncovers intricate patterns and dependencies. Neural networks are trained using backpropagation, an iterative procedure that repeatedly adjusts the network weights in the direction that minimizes the prediction error (Rumelhart, Hinton, & Williams, 1986). A central appeal of neural networks is their ability to approximate virtually any measurable function to an arbitrary degree of accuracy, providing a highly flexible framework for modeling complex relationships (Hornik, Stinchcombe, & White, 1989). This flexibility comes at a cost: neural networks generally require large datasets to train reliably, as well as substantial computational resources (Strittmatter, 2025).

While ML methods are excellent tools for modeling high-dimensional and nonlinear relationships, their foundation lies in prediction rather than causal inference. ML revolves around the prediction problem, producing predictions of Y from X. The appeal of ML is that it can uncover generalizable patterns and discover complex structures not specified in advance. By relying on regularization and out-of-sample evaluation, ML learns flexible functional forms that perform well for prediction without excessive overfitting. In contrast, many economic applications center on parameter estimation, producing a consistent estimate of the parameter θ that describes the causal effect of a treatment variable D on an outcome Y. Mullainathan & Spiess (2017) emphasize that ML algorithms are not built for this purpose. Even if ML methods produce coefficient estimates, they are rarely consistent, and interpreting them as causal effects may lead to misleading conclusions. Nevertheless, prediction has a long history in economics, where ML provides new tools to solve this old problem (Mullainathan & Spiess, 2017).

Examples of pure prediction problems in economics include forecasting default risk from large sets of financial characteristics, predicting consumer demand using high-dimensional product attributes, or forecasting stock market returns from macroeconomic indicators. These settings require flexible methods that can handle high-dimensional covariate spaces and nonlinear patterns, but the resulting predictions do not carry a causal interpretation. The focus lies solely on the accuracy of the predicted value, not on the causal effect of one variable on another (Mullainathan & Spiess, 2017).

1.3.2 From Predictive to Causal Machine Learning

Beyond pure prediction, a central application of ML is in causal inference, where estimation procedures rely on accurately predicting the outcome or treatment assignment (Mullainathan & Spiess, 2017). In high-dimensional settings, this requires selecting which control variables influence both the treatment and the outcome, since omitting relevant controls leads to omitted variable bias. At the same time, including all available controls can result in overfitting and imprecise estimates, particularly when the number of potential covariates is large relative to the sample size. ML tools offer a way to navigate this trade-off by selecting the relevant controls or by flexibly estimating the control function that captures the relationships between covariates, treatment, and outcome (Chernozhukov et al., 2018).

Belloni, Chernozhukov & Hansen (2014) propose a LASSO-based approach for settings where the number of potential covariates is large relative to the sample size. Their procedure, known as the post-double-selection method, applies LASSO separately to identify covariates that predict the treatment and covariates that predict the outcome. In the final step, the outcome is regressed on the treatment using the union of the variables selected in the previous two stages as controls. By ensuring that all covariates influencing either the treatment or the outcome are included, this method achieves consistent estimation in high-dimensional settings. Chernozhukov et al. (2018) generalize this idea in the double/debiased machine learning (DML) framework. DML uses flexible ML methods to predict treatment and outcome from the covariates, constructs residuals by subtracting these ML predictions, and then estimates the causal effect from the relationship between the residualized variables. By employing cross-fitting to obtain out-of-sample predictions of treatment and outcome, DML avoids overfitting and isolates the causal effect. This framework allows researchers to use other ML methods beyond LASSO, such as decision trees, random forests, and neural networks. These methods can help model complex confounding structures while retaining consistent and asymptotically normal estimates.

Using DML does not eliminate the need for a credible identification strategy. While DML effectively handles many potential confounders and accommodates flexible functional forms, its validity still depends on having access to all relevant confounders. If key confounders remain unobserved, DML can suffer from omitted variable bias just like traditional econometric approaches. Similarly, a careful selection of control variables is required to avoid collider bias, a problem to which DML is particularly sensitive. A collider bias occurs when conditioning on a variable that is influenced by both the treatment and the outcome, which distorts the estimated treatment effect. For instance, when estimating the effect of job training on employment, controlling for current earnings can generate collider bias. Job training affects earnings by improving skills and job prospects, while employment status also affects earnings. Conditioning on earnings may therefore create a spurious relationship in which individuals without training but with high earnings, perhaps due to prior experience or industry demand, appear more likely to be employed. This can make the impact of job training on employment seem negative, even though it is actually positive. DML is particularly vulnerable to this bias because colliders often have strong predictive power, and

ML models used within DML tend to rely heavily on such predictors. In contrast, traditional econometric methods, which place less emphasis on pure predictive accuracy, generally assign less weight to colliders. These methods still face the risk of collider bias but are typically less affected by it because colliders are not prioritized based on their predictive strength (Strittmatter, 2025).

A further contribution of ML to causal inference lies in uncovering heterogeneity in treatment effects. Traditional methods for analyzing heterogeneous treatment effects often rely on researchers' judgment in choosing variables relevant to heterogeneity. ML methods offer a systematic and data-driven alternative. A prominent example is the causal forest introduced by Wager & Athey (2018), which extends the random forest framework to estimate heterogeneous effects rather than to predict outcomes. The causal forest partitions the data using causal trees that split observations into subgroups based on differences in treatment responses. Unlike standard decision trees, which choose splits that minimize prediction errors, causal trees select splits that maximize differences in estimated treatment effects across leaves. Aggregating the results from many such trees yields estimates of conditional average treatment effects that can vary flexibly with the covariates. Similar to DML, causal forest estimates are consistent and asymptotically normal (Strittmatter, 2025).

2 Double/Debiased Machine Learning

2.1 Partially Linear Regression Model

To motivate the double/debiased machine learning (DML) approach, consider the partially linear regression (PLR) model as in Robinson (1988):

$$Y = D\theta_0 + g_0(X) + \zeta \quad \mathbb{E}[\zeta \mid D, X] = 0 \tag{2.1}$$

$$D = m_0(X) + V \quad \mathbb{E}[V \mid X] = 0 \tag{2.2}$$

where Y is the outcome variable and D is the treatment variable of interest. The vector $X = (X_1, \ldots, X_p)$ consists of confounding covariates, and ζ and V are stochastic errors. Equation (2.1) is the main regression equation, and θ_0 is the regression coefficient of interest. The confounders X affect the treatment variable D through the function $m_0(X)$ and the outcome variable Y through the function $g_0(X)$. If treatment D is randomly assigned conditional on covariates X, then θ_0 has the interpretation of the treatment effect parameter (Chernozhukov et al., 2018). Functions $m_0(X)$ and $g_0(X)$ have an unknown functional form, and misspecifying them would bias the estimate of θ_0, even if one controls for all covariates X. Figure 2.1 shows the causal diagram for the PLR under conditional exogeneity. Covariates X confound the identification of θ_0. The treatment effect θ_0 can be identified via the stochastic error V, which captures the exogenous variation in treatment D that does not depend on confounders X (Bach et al., 2024).

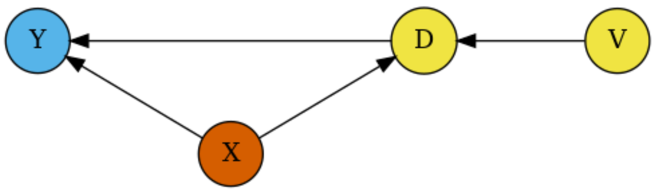

Figure 2.1 Causal diagram for PLR (Bach et al., 2024)

Chernozhukov et al. (2018) show that estimating the unknown function $g_0(X)$ naively using ML methods to control for the covariates X would bias the estimates of the treatment effect θ_0. This is because the ML estimators are generally biased, and can transmit this bias into θ_0. Nevertheless, using a flexible ML algorithm instead of specifying a parametric model for $g_0(X)$ is appealing, allowing researchers to relax assumptions about how the covariates enter the model. Flexibly adjusting for a large number of covariates increases the plausibility that all relevant confounding variation is considered (Belloni, Chernozhukov, Hansen, & Kozbur, 2016). To prepare the ground for the DML estimator, Chernozhukov et al. (2018) derive asymptotic properties of a naive estimator $\hat{\theta}_0$ that does not address the bias introduced by ML methods. Specifically, they split the sample into two subsamples: a main subsample of size n, with observation numbers indexed by

$i \in I$, and an auxiliary subsample of size $N - n$, with observations indexed by $i \in I^c$. They estimate $g_0(X)$ using the auxiliary subsample I^c and use its estimate to partial out the effect of confounders X from Y, yielding an estimator $\hat{\theta}_0$:

$$\hat{\theta}_0 = \left(\frac{1}{n} \sum_{i \in I} D_i^2 \right)^{-1} \frac{1}{n} \sum_{i \in I} D_i (Y_i - \hat{g}_0(X_i)) \tag{2.3}$$

where $\hat{g}_0(X_i)$ is the estimator of $g_0(X_i)$. Chernozhukov et al. (2018) show that the estimator $\hat{\theta}_0$ would be generally biased due to regularization bias introduced by the estimation of $g_0(X)$ using ML. ML regression models minimize the mean squared error $MSE(\hat{Y}) = \mathbb{E}[(Y - \hat{Y})^2]$, which decomposes into squared bias $(\mathbb{E}[\hat{Y}] - f(X))^2$, variance $\mathbb{E}[(\hat{Y} - \mathbb{E}[\hat{Y}])^2]$, and irreducible error σ_ε^2, where $Y = f(X) + \varepsilon$ and $Var(\varepsilon) = \sigma_\varepsilon^2$. Equation (2.4) illustratively summarizes the MSE decomposition.

$$\begin{aligned} MSE(\hat{Y}) &= \mathbb{E}[(Y - \hat{Y})^2] \\ &= \underbrace{(\mathbb{E}[\hat{Y}] - f(X))^2}_{bias^2} + \underbrace{\mathbb{E}[(\hat{Y} - \mathbb{E}[\hat{Y}])^2]}_{variance} + \underbrace{\sigma_\varepsilon^2}_{irreducible\ error} \end{aligned} \tag{2.4}$$

ML algorithms may introduce bias into the model to reduce variance, prevent overfitting, and improve the out-of-sample performance (Hastie, Tibshirani, & Friedman, 2009). To show the bias in the estimator $\hat{\theta}_0$ intuitively, I plug Y from (2.1) into (2.3) and examine the resulting probability limit. This approach complements the more technical derivation in Chernozhukov et al. (2018).

$$\hat{\theta}_0 = \theta_0 + \underbrace{\left(\frac{1}{n} \sum_{i \in I} D_i^2 \right)^{-1} \frac{1}{n} \sum_{i \in I} D_i \zeta_i}_{:=a} + \underbrace{\left(\frac{1}{n} \sum_{i \in I} D_i^2 \right)^{-1} \frac{1}{n} \sum_{i \in I} D_i (g_0(X_i) - \hat{g}_0(X_i))}_{:=b} \tag{2.5}$$

Equation (2.5) implies that term a converges in probability to zero by the law of large numbers and Slutsky's theorem. However, the term b does not vanish with growing sample size because it contains $\hat{g}_0(X_i)$ that suffers from regularization bias, transmitting it to the estimator $\hat{\theta}_0$.

2.2 Overcoming Regularization and Overfitting Biases

To overcome the regularization bias, Chernozhukov et al. (2018) construct a new estimator $\tilde{\theta}_0$, which they refer to as the DML estimator. Bach et al. (2024) rewrite the PLR model by Robinson (1988) in residualized form to make the derivation of the DML estimator by Chernozhukov et al. (2018) more transparent:

$$W = V\theta_0 + \zeta \quad \mathbb{E}[\zeta \mid D, X] = 0 \tag{2.6}$$

$$W = Y - \ell_0(X) \quad \ell_0(X) = \mathbb{E}[Y \mid X] \tag{2.7}$$

$$V = D - m_0(X) \quad m_0(X) = \mathbb{E}[D \mid X] \tag{2.8}$$

where the variables W and V represent the original variables Y and D, respectively, after partialling out the effect of confounders X. After applying this transformation,

W and V are orthogonal to any function of X. Note that V does not identify θ_0 if it has zero variance, meaning D cannot be entirely determined by X. The PLR model in residualized form captured by (2.6) – (2.8) yields the DML estimator $\tilde{\theta}_0$:

$$\tilde{\theta}_0 = \left(\frac{1}{n}\sum_{i\in I}\hat{V}_i^2\right)^{-1}\frac{1}{n}\sum_{i\in I}\hat{V}_i(Y_i - \hat{\ell}_0(X_i)) \tag{2.9}$$

where $\hat{V}_i = D_i - \hat{m}_0(X_i)$, $\hat{m}_0(X_i)$ and $\hat{\ell}_0(X_i)$ are estimators of $m_0(X_i)$ and $\ell_0(X_i)$, respectively. By partialling out the effect of X from D and Y, $\tilde{\theta}_0$ removes the regularization bias that contaminated the estimator $\hat{\theta}_0$ in Equation (2.3), hence the name debiased machine learning. DML performs two predictions using ML methods, the conditional mean of Y given X, denoted $\ell_0(X)$ and the conditional mean of D given X, denoted $m_0(X)$, hence the name double machine learning (Chernozhukov et al., 2018). To illustrate the benefits of orthogonalized variables and double predictions, I derive the probability limit of $\tilde{\theta}_0$, complementing the technical proofs by Chernozhukov et al. (2018).

$$\tilde{\theta}_0 = \theta_0 + \underbrace{\left(\frac{1}{n}\sum_{i\in I}\hat{V}_i^2\right)^{-1}\frac{1}{n}\sum_{i\in I}\hat{V}_i\zeta_i}_{:=a} + \underbrace{\left(\frac{1}{n}\sum_{i\in I}\hat{V}_i^2\right)^{-1}\frac{1}{n}\sum_{i\in I}\hat{V}_i(\ell_0(X_i) - \hat{\ell}_0(X_i))}_{:=b} \tag{2.10}$$

Equation (2.10) implies that term a converges in probability to zero, as in (2.5). In contrast, the term b now depends on the product of the estimation errors in $\ell_0(X)$ and $m_0(X)$. If these errors are independent, b vanishes in probability. However, if $\ell_0(X)$ and $m_0(X)$ are estimated using the same sample via highly complex ML methods, the estimation errors may be correlated, giving rise to the overfitting bias (Bach et al., 2024). To mitigate this concern, Chernozhukov et al. (2018) propose the cross-fitting procedure, where they estimate $\ell_0(X)$ and $m_0(X)$ on an auxiliary subsample I^c, and then evaluate the estimated functions on the main subsample I to form the out-of-sample predictions $\hat{\ell}_0(X)$ and $\hat{m}_0(X)$ used to construct $\tilde{\theta}_0$. This procedure prevents the ML models from overfitting on the main subsample, helping to ensure the prediction errors V_i and $\ell_0(X_i) - \hat{\ell}_0(X_i)$ are independent (Chernozhukov et al., 2018).

Chernozhukov et al. (2018) conduct simulations to illustrate the negative impacts of regularization and overfitting biases, highlighting the benefits of the DML estimator. In a PLR setup, defined by (2.1) – (2.2), they estimate $\ell_0(X)$ and $m_0(X)$ using random forests. Figure 2.2 shows the detrimental effect of regularization bias and the advantages of orthogonalizing Y and D. The left panel displays the simulated distribution of the non-orthogonal $\hat{\theta}_0 - \theta_0$ estimator, which is substantially biased and shifted noticeably to the right of zero. In contrast, the right panel depicts the distribution of the orthogonalized DML estimator, which is centered around zero and closely approximated by a normal distribution, suggesting that the bias is negligible. Figure 2.3 illustrates the adverse effect of an overfitting bias alongside the benefits of the cross-fitting procedure. The left panel shows that the simulated distribution for the orthogonalized $(\tilde{\theta}_0 - \theta_0)/se(\tilde{\theta}_0)$ estimator without cross-fitting is shifted markedly to the left, demonstrating the negative impact of overfitting bias. The right panel illustrates that this bias disappears after applying cross-fitting.

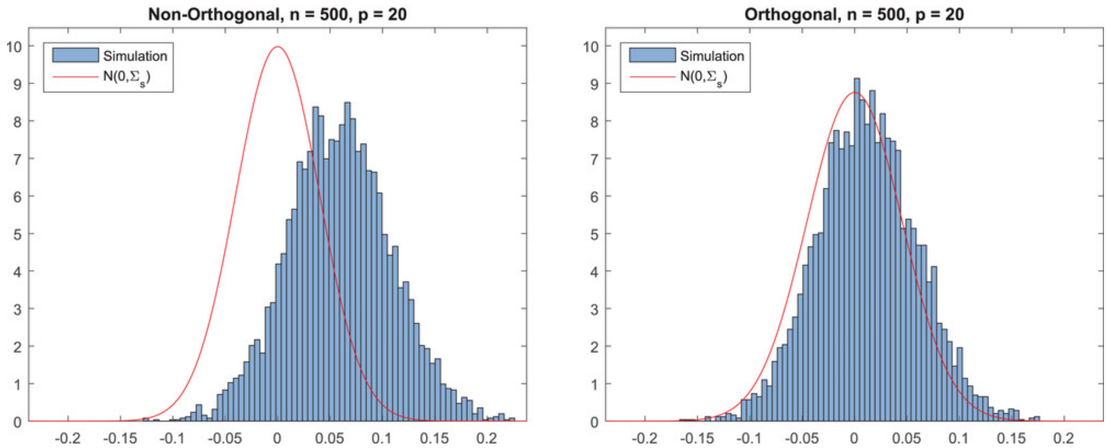

Figure 2.2 The effect of orthogonalization on the simulated distribution of $\hat{\theta}_0 - \theta_0$ estimator (Chernozhukov et al., 2018)

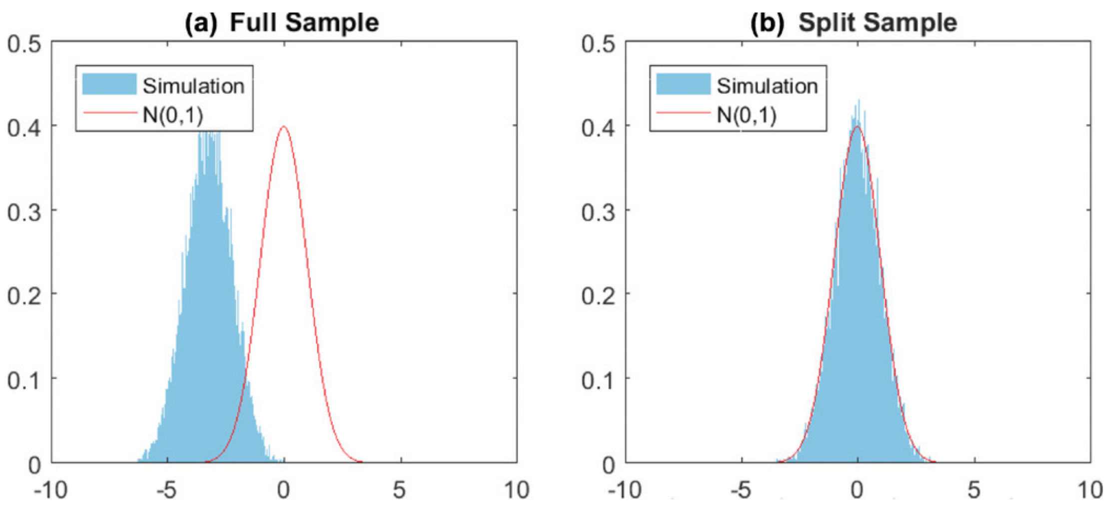

Figure 2.3 The effect of cross-fitting on the simulated distribution of $(\tilde{\theta}_0 - \theta_0)/se(\tilde{\theta}_0)$ estimator (Chernozhukov et al., 2018)

2.3 Neyman Orthogonality

Chernozhukov et al. (2018) generalize the orthogonalization principle captured in Equations (2.6) – (2.8) within the method-of-moments framework. They show that the parameter θ_0 can be estimated by solving the empirical analog of the moment condition:

$$\mathbb{E}[\psi(W;\theta_0,\eta_0)] = 0 \tag{2.11}$$

where ψ denotes the score function, $W = (Y, D, X)$, θ_0 is the population value of the parameter of interest, and $\eta = (\ell(X), m(X))$ marks the vector of nuisance functions with population value $\eta_0 = (\ell_0(X), m_0(X))$. To identify θ_0, the score function $\psi(W;\theta,\eta)$ must satisfy the Neyman orthogonality condition:

$$\partial_\eta \mathbb{E}[\psi(W;\theta_0,\eta)]|_{\eta=\eta_0} = 0 \tag{2.12}$$

where ∂_η denotes the directional derivative of the expected score function with respect to the nuisance parameter η. Neyman orthogonality (2.12) ensures that the moment condition (2.11) used to identify and estimate θ_0 is not sensitive to small perturbations of the nuisance function η around its true value η_0. This property allows for minor estimation errors in η, while still preserving consistent estimation of θ_0. Although estimating η_0 precisely is desirable, accurate predictions of η_0 do not necessarily lead to a consistent estimator of the treatment effect θ_0. Estimating η_0 using the ML estimator $\hat{\eta}_0$ could introduce a regularization bias to the estimation process, as I outlined in Equation (2.5). As a result, the estimator of θ_0 may cease to be unbiased and asymptotically normal. The Neyman-orthogonal score makes the estimator of θ_0 robust against first-order biases arising from regularization (Bach et al., 2024).

Chernozhukov et al. (2018) show that the PLR summarized by (2.1) – (2.2) uses a score function:

$$\psi(W;\theta,g(X)) = (Y - \theta D - g(X))D \tag{2.13}$$

which does not satisfy the Neyman orthogonality defined in (2.12). This is because the score function $\psi(W;\theta,g(X))$ is sensitive to small estimation errors in $g(X)$, such as those introduced by regularization, and these errors contaminate the estimation of θ_0. This result is consistent with my derivation of the probability limit of $\hat{\theta}_0$ estimator in (2.5). By contrast, the PLR model in residualized form captured by (2.6) – (2.8) uses a score function:

$$\psi(W;\theta,\eta) = (Y - \ell(X) - \theta(D - m(X)))(D - m(X)) \tag{2.14}$$

which satisfies the Neyman orthogonality principle (Bach et al., 2024). I interpret Equation (2.14) within the method-of-moments framework as the moment condition formed by the product of the regression error and the residualized treatment variable. If their product is zero in the expectation, regression produces consistent estimates. Equation (2.14) shows that even if $\ell(X)$ and $m(X)$ are estimated with small errors that are random, these errors cancel out in expectation, satisfying $\mathbb{E}[\psi(W;\theta,\eta)] = 0$. However, if the estimation errors in $\ell(X)$ and $m(X)$ are systematic, the residuals may become correlated due to overfitting. This leads to the violation of the moment condition (2.11) and inconsistency of $\tilde{\theta}_0$, in line with my derivation in Equation (2.10).

Bach et al. (2024) summarize the cross-fitting procedure introduced by Chernozhukov et al. (2018) to mitigate the bias induced by overfitting when using ML for estimating nuisance parameters. While Chernozhukov et al. (2018) describe two similar algorithms for cross-fitting, referred to as DML1 and DML2, they recommend using DML2 for the more stable behavior of the estimates. Following Bach et al. (2024), I outline the DML2 algorithm in Definition (1).

Definition 1 (DML2 Algorithm)**.**

(1) Provide the observed data $(W_i)_{i=1}^N$, a Neyman-orthogonal score function $\psi(W;\theta,\eta)$ and specify ML models to estimate η.

(2) Partition the sample of size N randomly into K equally sized folds $(I_k)_{k=1}^K$, so that each fold I_k has a size N/K. For each fold $k \in [K] = \{1,\dots,K\}$, construct a high-quality ML estimator of η:

$$\hat{\eta}_{0,k} = \hat{\eta}_{0,k}\left((W_i)_{i \notin I_k}\right) \tag{2.15}$$

where the nuisance function estimator for fold k, $\hat{\eta}_{0,k}$, is estimated on the remaining $K-1$ folds not in I_k and fitted out-of-sample on the evaluation set I_k.

(3) Construct the estimator for the causal parameter $\tilde{\theta}_0$ as the solution to the equation:

$$\frac{1}{N}\sum_{k=1}^{K}\sum_{i \in I_k}\psi(W_i;\tilde{\theta}_0,\hat{\eta}_{0,k}) = 0 \tag{2.16}$$

Definition (1) emphasizes the importance of using high-quality ML models for η_0. This is because the Neyman orthogonality ensures that the estimator $\tilde{\theta}_0$ remains consistent only if the estimation error in η_0 is relatively small. Chernozhukov et al. (2018) show that the product of the estimation error in m_0 and the sum of the errors in m_0 and ℓ_0 must converge to zero faster than $N^{-1/2}$ as the sample size N increases:

$$\|\hat{m}_0 - m_0\|_{P,2} \times (\|\hat{m}_0 - m_0\|_{P,2} + \|\hat{\ell}_0 - \ell_0\|_{P,2}) \leq o(N^{-1/2}) \tag{2.17}$$

where $\|\cdot\|_{P,2}$ denotes the L^2 norm under the data-generating process P. Equation (2.17) highlights that in regression settings, minimizing the out-of-sample root mean squared error (RMSE) of the nuisance function estimators $\hat{\ell}_0$ and $\hat{m}_0$ should result in more precise estimates of θ_0. This is because the DML procedure relies on L^2-convergence rates of the nuisance functions for valid inference (Chernozhukov et al., 2018).

Chernozhukov et al. (2018) recommend various ML tools to estimate η_0 based on different structural assumptions of η_0:

(1) For sparse structures of η with respect to some set of regressors, they recommend sparsity-based ML methods, such as lasso, post-lasso, l_2-boosting, or forward selection.

(2) For dense structures of η with respect to some set of regressors, they suggest density-based estimators such as ridge. For mixed structures based on sparsity and density, they suggest using the elastic net or lava.

(3) If tree-based methods can capture the structure of η_0, they advise using regression trees or random forest.

(4) When either sparse, shallow, or deep neural networks can approximate the structure of η_0, they recommend l_1-penalized neural networks, shallow neural networks, or deep neural networks.

2.4 Inference

Chernozhukov et al. (2018) derived asymptotic properties of the DML estimator $\tilde{\theta}_0$ in the PLR model. Following their results, I summarize these properties in Theorem (1), which I adapted from Bach et al. (2024).

Theorem 1 (DML Inference in PLR)**.** *Under suitable regularity conditions, the DML estimator $\tilde{\theta}_0$ is approximately normal:*

$$\sqrt{N}(\tilde{\theta}_0 - \theta_0) \rightsquigarrow N(0, \sigma^2) \tag{2.18}$$

with zero mean and asymptotic variance given by:

$$\sigma^2 = J_0^{-2}\mathbb{E}[\psi^2(W; \theta_0, \eta_0)] \tag{2.19}$$

$$J_0 = \mathbb{E}[\psi_a(W; \eta_0)] \tag{2.20}$$

where $J_0 = \mathbb{E}[\psi_a(W; \eta_0)] = \mathbb{E}[\partial_\theta \psi(W; \theta, \eta_0)]$ is the derivative of the moment condition with respect to the parameter of interest θ and $\mathbb{E}[\psi^2(W; \theta_0, \eta_0)] = \mathrm{Var}[\psi(W; \theta_0, \eta_0)]$ is its variance.

Theorem (1) shows that, when the Neyman-orthogonal score and cross-fitting are used, the DML estimator has well-behaved sampling properties, so inference can be carried out using the usual normal approximation. The empirical analogues of the asymptotic variance components given by (2.19) and (2.20) are:

$$\hat{\sigma}^2 = \hat{J}_0^{-2} \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in I_k} \left[\psi(W_i; \tilde{\theta}_0, \hat{\eta}_{0,k})\right]^2 \tag{2.21}$$

$$\hat{J}_0 = \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in I_k} \psi_a(W_i; \hat{\eta}_{0,k}) \tag{2.22}$$

The asymptotic variance estimator $\hat{\sigma}^2$ can be used to construct a $(1-\alpha)$ confidence interval for $\tilde{\theta}_0$ as:

$$\left[\tilde{\theta}_0 \pm q_{1-\alpha/2}^{N(0,1)} \frac{\hat{\sigma}}{\sqrt{N}}\right] \tag{2.23}$$

where $q_{1-\alpha/2}^{N(0,1)}$ is an $(1-\alpha/2)$-quantile of the standard normal distribution (Bach et al., 2024).

2.5 Heterogeneity Analysis

DML methodology allows treatment effects θ_0 to vary for different groups by slightly modifying the PLR model. To accommodate heterogeneity, the constant treatment effect is replaced by a function of observed characteristics:

$$Y = D\theta_0(X) + g_0(X) + \zeta \quad \mathbb{E}[\zeta \mid D, X] = 0 \tag{2.24}$$

$$D = m_0(X) + V \quad \mathbb{E}[V \mid X] = 0 \tag{2.25}$$

where $\theta_0(X)$ represents the CATE function. It holds:

$$\mathbb{E}[Y \mid X] = \mathbb{E}[\theta_0(X)D \mid X] + \mathbb{E}[g_0(X) \mid X] + \underbrace{\mathbb{E}[\zeta \mid X]}_{=\mathbb{E}[\mathbb{E}[\zeta|D,X]|X]=0} \tag{2.26}$$

$$= \theta_0(X)\mathbb{E}[D \mid X] + g_0(X) \tag{2.27}$$

such that:

$$\underbrace{Y - \mathbb{E}[Y \mid X]}_{:=W} = \theta_0(X)\underbrace{(D - \mathbb{E}[D \mid X])}_{:=V} + \zeta \tag{2.28}$$

By the law of iterated expectations:

$$\mathbb{E}[\zeta \mid V] = \mathbb{E}[\mathbb{E}[\zeta \mid X, D] \mid V] = 0 \tag{2.29}$$

and consequently:

$$\mathbb{E}[W \mid V] = \theta_0(X)V \tag{2.30}$$

Therefore, $\theta_0(X)$ can be estimated by regressing W on V:

$$\theta_0(X) = \arg\min_{\theta(X)\in\Theta} \mathbb{E}\left[(W - \theta(X)V)^2\right] \tag{2.31}$$

When θ_0 depends on a discrete grouping variable G_k, rather than a continuous variable, the problem simplifies considerably. In this case, CATE can be modeled as group-specific coefficients, and estimation reduces to interacting the residualized treatment V with the group indicators G_k. The objective function in Equation (2.31) therefore becomes:

$$\theta_k = \arg\min_{\theta\in\Theta} \mathbb{E}\left[(W - \theta\; VG_k)^2\right] \tag{2.32}$$

where θ_k is the CATE for group k. This expression highlights that, once the outcome and treatment have been residualized using ML, the group-specific CATEs can be obtained from a linear regression of the residualized outcome on the interaction between the residualized treatment and the group indicator (Bach et al., 2024).

3 Empirical Application

3.1 Empirical Setting

My thesis builds upon the paper by Reimers & Waldfogel (2021), which estimates the aggregate effects of price, Amazon star ratings, and newspaper reviews on consumer demand in the Amazon book marketplace. The authors then use the estimated demand elasticities to calibrate the nested logit model of demand à la Berry (1994) to study how star ratings and newspaper reviews influence consumer welfare. Specifically, they employ the linear FE model to determine the causal effect of price, star ratings, and newspaper reviews on Amazon sales rank:

$$\ln(rank_{jct}) = \theta_1 \ln(rank_{jc,t-1}) + \theta_2 \ln(p_{jct}) + \theta_3 \ln(stars_{jct}) + \theta_4 \ln(ratings_{jct}) + \theta_5 \ln(stars_{jct}) \cdot \ln(ratings_{jct}) + h_{\tau c} + f(U_{jt}, S_{jt}) + \mu_{jc} + \zeta_{jct} \quad (3.1)$$

where $rank_{jct}$ is the sales rank, p_{jct} is the price, $stars_{jct}$ is the Amazon star rating, $ratings_{jct}$ is the number of customer ratings for book title j on platform c on day t. Chevalier & Goolsbee (2003) shows that sales rank is a reasonable proxy for demand because the relationship between the natural logarithm of sales rank and the natural logarithm of demand volume is approximately linear. The term $h_{\tau c}$ is a dummy variable equal to one on each platform c for τ days after the appearance of a professional review in a newspaper, capturing the differential impact of newspaper reviews across Amazon platforms $c \in \{\text{United States}, \text{Canada}, \text{United Kingdom}\}$. To estimate the dynamic effect of professional reviews, the authors construct a set of time-specific indicators based on the number of days following the appearance of a review. For reviews published in the New York Times (NYT), they define three time intervals $\tau \in \{\text{0-5 days}, \text{6-10 days}, \text{11-20 days}\}$ after the review is published. For the other major U.S. newspapers, including the Boston Globe, the Chicago Tribune, the Los Angeles Times, the Wall Street Journal, and the Washington Post, the authors define two intervals $\tau \in \{\text{0-10 days}, \text{11-20 days}\}$. In addition, they distinguish between books that the NYT merely reviewed and those that it explicitly recommended, allowing for different effects based on the positivity of a review. In total, they estimate 24 dummy variables: 8 per platform, corresponding to the NYT-reviewed and NYT-recommended books across the three post-review intervals and non-NYT reviews across two intervals. Nevertheless, they report the effects of newspaper reviews only for the U.S. platform, using the effects on the British and Canadian platforms as control variables. U_{jt} is the number of days until the publication of a title j on the day t, with zero values after the publication. S_{jt} is the number of days since the publication of a title j on the day t, with zero values before the publication. $f(U_{jt}, S_{jt})$ includes first- through third-order polynomial terms of U_{jt} and S_{jt} to flexibly account for time patterns of sales before and after the publication date. Note that Amazon allows pre-ordering of books before they are officially published. μ_{jc} are platform-edition FE controlling for time-invariant unobserved heterogeneity of each book edition on each platform (Reimers & Waldfogel, 2021).

Reimers & Waldfogel (2021) emphasize that star ratings and newspaper reviews are inherently endogenous. Raters and reviewers decide whether and when to give feedback in addition to what they write. This creates reverse causality as more appealing books sell more and receive more positive feedback. The authors employ two strategies to address endogeneity, one for newspaper reviews and another for star ratings. They treat the appearance of a professional review as a discontinuous jump in attention delivered to the title and look for a corresponding jump in daily sales. By examining sales in a short period after the review appears, they isolate its effect from other factors that might gradually influence sales over time. To overcome the endogeneity of star ratings, Reimers & Waldfogel (2021) measure their impact following Chevalier & Mayzlin (2006), employing platform-edition FE to control for time-invariant unobserved heterogeneity of each book edition on each platform. Tracking the same book across multiple platforms over time offers them a natural control for each book. When a given book exhibits different star ratings and sales across platforms, controlling for platform-edition FE isolates the causal impact of star ratings on sales (Chevalier & Mayzlin, 2006).

Although adequate for welfare analysis, the constant ATE assumption of price, star ratings, and newspaper reviews may seem unrealistic given the thousands of book editions, authors, and publishers in the Amazon book marketplace. Understanding market heterogeneity is crucial for booksellers and marketers, as it enables them to tailor differentiated product offerings to each market segment. Although it is likely that each book would respond differently to shifts in price, star ratings, or newspaper reviews, I aim to analyze heterogeneity across book genres because they are a universally understood category that points to distinct reader segments. Everyone understands what to expect from a horror versus a romance, while my results could be harder to grasp if my analysis centers around authors or titles that the general population is unfamiliar with. By analyzing genres, I can draw more generalizable insights into demand heterogeneity.

3.2 Data

To investigate heterogeneity in consumer demand across book genres in the Amazon book marketplace, I use a high-frequency panel of book demand by Reimers & Waldfogel (2021). It includes daily data from Amazon domains in the United States, Canada, and the United Kingdom, covering Amazon sales ranks, prices, Amazon star ratings, the number of customer ratings, the number of days until publication, and the number of days since publication for 13,645 book editions sold in 2018. I use sales rank as a proxy for consumer demand, where a low sales rank indicates high demand and a high sales rank indicates low demand (Chevalier & Goolsbee, 2003). The panel also contains information on whether major U.S. newspapers reviewed the books, including the NYT, the Boston Globe, the Chicago Tribune, the Los Angeles Times, the Wall Street Journal, and the Washington Post. After dropping observations with missing values for any of the sales rank, price, star rating, number of customer ratings, newspaper review, number of days until publication, or number of days since publication, the resulting dataset contains 4,210,233 observations.

Genre	# Editions	Mean rank	Mean price	Mean star rating	Mean reviews	# Reviewed Editions
ART	136	188,951	19.93	4.36	12	42
BIOGRAPHY & AUTOBIOGRAPHY	854	144,183	16.97	4.35	40	209
BUSINESS & ECONOMICS	143	37,631	18.72	4.38	31	34
COMICS & GRAPHIC NOVELS	496	215,256	16.39	4.32	12	9
COOKING	147	16,920	18.50	4.45	25	4
DRAMA	80	367,454	11.74	4.25	60	3
EDUCATION	121	105,507	15.78	4.35	22	23
FAMILY & RELATIONSHIPS	83	50,843	14.67	4.53	64	28
FICTION: Action & Adventure	80	163,276	14.51	4.36	113	3
FICTION: Biographical	84	139,073	16.29	4.25	95	15
FICTION: Coming of Age	122	125,372	16.27	4.13	111	30
FICTION: Crime	230	170,014	14.60	4.13	64	15
FICTION: Cultural Heritage	171	95,610	16.27	4.32	158	20
FICTION: Dystopian	61	116,345	15.68	4.29	477	9
FICTION: Family Life	505	154,917	16.12	4.19	118	71
FICTION: Fantasy	304	135,243	15.76	4.37	126	23
FICTION: Friendship	76	220,636	15.77	4.14	153	10
FICTION: General	583	239,001	14.45	4.19	60	53
FICTION: Historical	352	134,215	16.01	4.32	108	66
FICTION: Horror	100	189,215	14.78	4.19	100	5
FICTION: Humorous	79	137,551	16.12	4.14	96	21
FICTION: Literary	104	254,399	15.20	4.17	107	6
FICTION: Mystery & Detective	784	217,850	14.89	4.26	67	26
FICTION: Psychological	178	148,624	15.18	4.06	95	40
FICTION: Religious	72	191,541	13.66	4.42	35	7
FICTION: Romance	911	223,074	11.79	4.34	64	25
FICTION: Science Fiction	217	150,541	14.52	4.21	91	9
FICTION: Thrillers	694	117,004	14.38	4.18	89	51
FICTION: Women	97	82,091	17.12	3.99	41	41
HEALTH & FITNESS	92	38,001	16.86	4.45	47	8
HISTORY	271	120,390	18.70	4.33	24	92
HOUSE & HOME	110	27,674	15.43	4.33	39	14
HUMOR	54	861,98	15.02	4.39	19	12
JUVENILE FICTION	2,437	166,375	11.34	4.49	34	69
JUVENILE NONFICTION	354	148,942	11.89	4.45	20	9
LITERARY COLLECTIONS	135	194,275	15.99	4.19	15	71
LITERARY CRITICISM	254	625,030	17.95	4.05	12	43
MEDICAL	52	25,728	17.81	4.29	34	17
MUSIC	25	259,177	15.18	3.47	6	6
NATURE	52	99,445	16.83	4.54	11	20
PHILOSOPHY	30	36,135	17.59	4.31	19	14
POETRY	132	123,392	14.95	4.44	11	50
POLITICAL SCIENCE	172	35,236	19.46	4.33	19	100
PSYCHOLOGY	106	29,325	17.06	4.35	64	30
RELIGION	158	67,555	14.87	4.45	22	16
SCIENCE	162	71,352	17.19	4.37	23	64
SOCIAL SCIENCE	208	63,540	17.87	4.39	21	93
SPORTS & RECREATION	59	56,380	19.23	4.40	13	17
TRAVEL	29	61,045	16.94	4.38	19	7
UNKNOWN	889	343,707	12.42	4.41	21	20
TOTAL	13,645	146,675	14.46	4.33	43	1670

Table 3.1 Descriptive statistics of book editions by genre

Table 3.1 presents descriptive statistics of book editions by genre, as genres are central to my analysis. For each genre, the table reports the number of editions in that genre, the average sales rank, price, star rating, and number of customer ratings across all editions belonging to that genre, as well as the number of editions reviewed by newspaper outlets in that genre. The final row provides overall sums or averages across all genres for comparison.

The table shows that among the 50 genres in my analysis, Juvenile Fiction is the most common with a total of 2,437 editions, followed by Fiction: Romance with 911 editions. Genres with the fewest editions are Music, with 25 editions, and Travel, with 29 editions. The best-selling genre is Cooking, with a mean sales rank of 16,920, followed by Medical with a mean rank of 25,728. The worst-selling genres are Literary Criticism, with a rank of 625,030, and Drama, with a mean rank of 367,454. The cheapest genre is Juvenile Fiction with an average price of $11.34, followed by Drama with a mean price of $11.74. The priciest genres include Art with an average price of $19.93 and Political Science with a mean price of $19.46. The highest-rated genre is Nature, with a mean star rating of 4.54, followed by Family & Relationships with a mean rating of 4.53. The lowest-rated genres are Music, with a mean star rating of 3.47, and Fiction: Women, with a mean rating of 3.99. The genre with the most customer ratings is Fiction: Dystopian, with an average of 477 consumer ratings, followed by Fiction: Cultural Heritage with an average of 158 ratings. The genres with the fewest customer ratings are Music, with an average of 6 ratings, and Nature, with an average of 11 ratings. Political Science is the most reviewed genre by review outlets, with 100 out of 172 editions reviewed (about 58%), followed by Literary Collections, with 71 out of 135 editions reviewed (about 53%). The genres with the lowest share of reviewed editions are Comics & Graphic Novels, with 9 out of 496 editions reviewed (about 2%), and Juvenile Nonfiction, with 9 out of 354 editions reviewed (about 3%).

Note that the dataset by Reimers & Waldfogel (2021) contains 126 genre labels, many of which are truncated, duplicated, or extremely small. Because the heterogeneity analysis requires sufficiently large groups, I merged these labels into 50 genres so that each genre contains at least 25 editions. This included correcting misspellings, merging duplicate categories, and merging very small genres into closely related larger ones. Table A.4 in the Appendix shows representative examples of these merging rules.

4 Empirical Strategies

To make heterogeneous treatment effects across genres interpretable, I pool the effects of newspaper reviews $h_{\tau c}$ for all Amazon platforms c into one variable $news_{jt}$. This includes the book edition dummies capturing the post-review effect for 0-5 days, 6-10 days, and 11-20 days after the review that the NYT either reviewed or recommended, and the post-review dummies for 0-10 days and 11-20 days that other major U.S. newspapers reviewed. The new variable $news_{jt}$ equals one for a book title j during the 20-day window following a professional newspaper review and zero otherwise. In total, $news_{jt}$ aggregates the 24 dummies of Reimers & Waldfogel (2021), 8 for each Amazon platform. While useful for estimating short-term impacts, these timing-specific effects are overly detailed for analyzing broader heterogeneity across genres. Aggregating the post-review effects into one variable facilitates the interpretability of heterogeneous effects that would otherwise involve interacting 50 genre dummies with 24 newspaper review dummies in specification (3.1).

4.1 Linear Fixed-Effects Model

To empirically investigate the demand elasticities across book genres, I employ the CATE framework and extend a linear FE model in (3.1) by interacting the treatment variables $\ln(p_{jct})$, $\ln(stars_{jct})$, and $news_{jt}$ with genre dummies:

$$\begin{aligned}\ln(rank_{jct}) = {} & \Theta_1 \ln(p_{jct}) \cdot genre_i + \Theta_2 \ln(stars_{jct}) \cdot genre_i + \Theta_3 news_{jt} \cdot genre_i \\ & + \theta_4 \ln(rank_{jc,t-1}) + \theta_5 \ln(reviews_{jct}) + f(U_{jt}, S_{jt}) + \mu_{jc} + \zeta_{jct} \quad (4.1)\end{aligned}$$

where $genre_i$ is a genre indicator that equals one if title j belongs to genre i and zero otherwise. Θ_1, Θ_2, and Θ_3 are vectors of coefficients, each of dimension 50×1, measuring genre-specific impacts of changes in price, star rating, and newspaper reviews, respectively.

4.2 Double/Debiased Machine Learning

While easy to interpret, the linear FE model with dummy interactions in (4.1) may not fully capture the complex relationships of consumer demand. The variables may interact nonlinearly, which the linear model does not account for. Notably, Reimers & Waldfogel (2021) suggest that relationships between variables may be nonlinear by including first- through third-order polynomial terms of the number of days until and since the publication $f(U_{jt}, S_{jt})$ in their specification (3.1). Following Reimers & Waldfogel (2021), I account for nonlinear effects of U_{jt} and S_{jt} using the polynomial terms, but if additional nonlinearities are present, my CATE estimates would be biased. Wooldridge (2020) stresses that the estimates in the linear regression model would be biased even if all relevant variables are included, but in the wrong functional form. A tolerance for functional form errors shrinks when estimating CATE. A model that yields unbiased ATEs can still give biased CATEs. This is because the treatment-genre interactions may pick up genre-specific nonlinearities that often average out in a pooled regression to estimate ATE (Blackwell & Olson, 2022). I estimate CATEs for price, star ratings, and newspaper reviews in my setting, resulting in a total of 150 CATE coefficients. With so many subgroup slopes, the risk of having at least some distorted by unmodeled nonlinearities is far higher than when estimating a single ATE. To avoid misspecifying the functional form by assuming pure linear relationships, I employ the DML approach by Chernozhukov et al. (2018), which allows accounting for the nonlinear effects of confounding variables. The DML model constitutes the main specification of this thesis. I will compare its estimates with those from the linear FE model, which serves as a benchmark to examine whether nonlinearities bias my genre-specific elasticity estimates.

Specifically, I define the following system with genre-specific treatment effects that I want to estimate:

$$\begin{aligned}\ln(rank_{jct}) = {} & \Theta_1 \ln(p_{jct}) \cdot genre_i + \Theta_2 \ln(stars_{jct}) \cdot genre_i + \Theta_3 news_{jt} \cdot genre_i \\ & + g(\mathbf{X_{jct}}) + \mu_{jc} + \zeta_{jct}\end{aligned} \tag{4.2}$$

where $\mathbf{X_{jct}} = (\ln(rank_{jc,t-1}), \ln(reviews_{jct}), U_{jt}, S_{jt})$. Equation (4.2) represents the underlying structural relationship that I aim to identify, linking book demand to price, star ratings, and newspaper reviews while allowing their effects to vary by genre. Rather than estimating this full equation directly, I recover its parameters Θ_1, Θ_2, and Θ_3 using the DML framework, which provides an algebraically equivalent estimation based on residualized variables.

I first remove the FE μ_{jc} using the within-group (WG) transformation, denoting WG-transformed variables Z with a tilde $\widetilde{Z}$. Fuhr & Papies (2024) caution against applying the WG transformation in nonlinear models, as its equivalence to the linear case no longer holds. In a nonlinear setting, the correct within transformation would require subtracting $\bar{g}(X_{jc})$, but $\bar{g}(X_{jc})$ is unknown. In practice, this is approximated by applying the WG transformation to the covariates X_{jct} and letting the ML algorithm learn $g(X_{jct} - \bar{X}_{jc})$, which is not equivalent to $g(X_{jct}) - \bar{g}(X_{jc})$. This approximation becomes problematic when the time dimension T is small, because the demeaned covariates $X_{jct} - \bar{X}_{jc}$ then contain limited within-unit variation, making it difficult for the ML algorithm to recover the nonlinear mapping. However, Fuhr & Papies (2024) show that when T is large, the within-unit variation after demeaning remains rich enough for the algorithm to learn even complex nonlinear effects. In my dataset, each book-platform pair has, on average, 167 daily observations, providing sufficient within-unit variability to model the nonlinear relationship reliably even after applying the WG transformation.

I then implement the DML procedure separately for each treatment variable $\widetilde{\ln}(p_{jct})$, $\widetilde{\ln}(stars_{jct})$, and $\widetilde{news}_{jt}$. Each DML estimation isolates the causal effect of the specific treatment by residualizing both the outcome and the treatment with respect to the full set of confounders $\widetilde{X}_{jct}$, which also includes the other treatments not currently analyzed. This procedure is algebraically equivalent to estimating the full model in Equation (4.2), as it mirrors the multi-treatment DML estimation framework and the implementation described by Bach et al. (2024) in Python's `DoubleML` package. I adopt this sequential single-treatment approach because the current implementation in Python's `DoubleML` package supports CATE estimation only for one treatment variable at a time. Therefore, I replicate the single-treatment DML estimation sequentially for each of $\widetilde{\ln}(p_{jct})$, $\widetilde{\ln}(stars_{jct})$, and $\widetilde{news}_{jt}$ to obtain genre-specific effects. Finally, I recover genre-specific elasticities by regressing the residualized outcome on the interaction between the residualized treatment and the genre dummies $genre_i$ as outlined in Equation 2.32.

I illustrate the estimation procedure for each treatment below. For the price treatment:

$$\widetilde{\ln}(rank_{jct}) = \theta_1 \widetilde{\ln}(p_{jct}) + g_1(\widetilde{\mathbf{X}}_{\mathbf{1,jct}}) + \widetilde{\zeta}_{1,jct} \tag{4.3}$$

$$\widetilde{\ln}(p_{jct}) = m_1(\widetilde{\mathbf{X}}_{\mathbf{1,jct}}) + \widetilde{V}_{1,jct} \tag{4.4}$$

where $\widetilde{\mathbf{X}}_{\mathbf{1,jct}} = (\widetilde{\ln}(rank_{jc,t-1}), \widetilde{\ln}(reviews_{jct}), \widetilde{U}_{jt}, \widetilde{S}_{jt}, \widetilde{\ln}(stars_{jct}), \widetilde{news}_{jt})$.

For the star ratings treatment:

$$\widetilde{\ln}(rank_{jct}) = \theta_2 \widetilde{\ln}(stars_{jct}) + g_2(\widetilde{\mathbf{X}}_{\mathbf{2,jct}}) + \widetilde{\zeta}_{2,jct} \tag{4.5}$$

$$\widetilde{\ln}(stars_{jct}) = m_2(\widetilde{\mathbf{X}}_{\mathbf{2,jct}}) + \widetilde{V}_{2,jct} \tag{4.6}$$

where $\widetilde{\mathbf{X}}_{\mathbf{2,jct}} = (\widetilde{\ln}(rank_{jc,t-1}), \widetilde{\ln}(reviews_{jct}), \widetilde{U}_{jt}, \widetilde{S}_{jt}, \widetilde{\ln}(p_{jct}), \widetilde{news}_{jt})$.

For the newspaper reviews treatment:

$$\widetilde{\ln}(rank_{jct}) = \theta_3 \widetilde{news}_{jt} + g_3(\widetilde{\mathbf{X}}_{\mathbf{3,jct}}) + \widetilde{\zeta}_{3,jct} \tag{4.7}$$

$$\widetilde{news}_{jt} = m_3(\widetilde{\mathbf{X}}_{\mathbf{3,jct}}) + \widetilde{V}_{3,jct} \tag{4.8}$$

where $\widetilde{\mathbf{X}}_{\mathbf{3,jct}} = (\widetilde{\ln}(rank_{jc,t-1}), \widetilde{\ln}(reviews_{jct}), \widetilde{U}_{jt}, \widetilde{S}_{jt}, \widetilde{\ln}(p_{jct}), \widetilde{\ln}(stars_{jct}))$.

In the final step, I estimate genre-specific elasticities for each treatment from the residualized variables via a linear regression:

$$\widetilde{W}_{k,jct} = \Theta_k \widetilde{V}_{k,jct} \cdot genre_i + \widetilde{\zeta}_{k,jct} \tag{4.9}$$

where $\widetilde{W}_{k,jct} = \widetilde{\ln}(rank_{jct}) - \ell_k(\widetilde{\mathbf{X}}_{\mathbf{k,jct}})$ denotes the residualized outcome, and $\widetilde{V}_{k,jct} = \widetilde{D}_{k,jct} - m_k(\widetilde{\mathbf{X}}_{\mathbf{k,jct}})$ is the residualized treatment. I estimate the nuisance functions $\ell_k(\widetilde{\mathbf{X}}_{\mathbf{k,jct}}) = \mathbb{E}[\widetilde{\ln}(rank_{jct}) \mid \widetilde{\mathbf{X}}_{\mathbf{k,jct}}]$ and $m_k(\widetilde{\mathbf{X}}_{\mathbf{k,jct}}) = \mathbb{E}[\widetilde{D}_{k,jct} \mid \widetilde{\mathbf{X}}_{\mathbf{k,jct}}]$ using a neural network. The variable $\widetilde{D}_{k,jct}$ denotes the treatment variable, where $\widetilde{D}_{1,jct} = \widetilde{\ln}(p_{jct})$ for price, $\widetilde{D}_{2,jct} = \widetilde{\ln}(stars_{jct})$ for star ratings, and $\widetilde{D}_{3,jct} = \widetilde{news}_{jt}$ for newspaper reviews.

4.2.1 Neural Network Architecture

I estimate the nuisance functions $\ell_k(\widetilde{\mathbf{X}}_{\mathbf{k,jct}})$ and $m_k(\widetilde{\mathbf{X}}_{\mathbf{k,jct}})$ using a multilayer perceptron, which is a fully connected feedforward neural network. I use this neural network architecture because it can approximate any measurable function to any desired degree of accuracy, allowing it to flexibly capture the nonlinear relationships between confounders and the treatment, as well as between confounders and the outcome (Hornik, Stinchcombe, & White, 1989). I apply five-fold cross-fitting, meaning the nuisance functions for each held-out fold are trained using the remaining four folds. Furthermore, the four folds used for neural network training contain 3,368,186 observations (out of a total of 4,210,233), which should provide sufficient data points to reliably learn the underlying relationships.

The neural network consists of four hidden layers with 64 neurons each, a hyperbolic tangent activation function, and L^2 regularization of 10^{-4} on all hidden-layer weights. The input layer has dimension six, corresponding to the four covariates $\widetilde{\ln}(rank_{jc,t-1})$, $\widetilde{\ln}(reviews_{jct})$, $\widetilde{U}_{jt}$, and $\widetilde{S}_{jt}$, together with the two treatments for which CATEs are not currently being estimated. The output layer is linear. I use the hyperbolic tangent activation function because, in simple regression problems on tabular data, it often performs better than more modern activation functions such as the rectified linear unit (ReLU) by preserving both positive and negative activations, which can carry meaningful information in economic applications (Szandała, 2021). In a neural network, each neuron computes a weighted sum of its inputs, and the activation function determines whether this signal is passed forward as a positive or negative value to the next layer. This property is particularly relevant in economic settings, where covariates are often WG-transformed and negative values naturally indicate below-average realizations, so discarding them would remove relevant information. In contrast, ReLU sets all negative activations to zero and is typically more effective in image and sound recognition tasks, where threshold-based activations are advantageous and negative values generally carry little information for pattern recognition.

I standardize all covariates before fitting, so they have a mean of zero and a variance of one. I train the neural network via backpropagation (Rumelhart, Hinton, & Williams, 1986) using the Adam optimizer (Kingma & Ba, 2014) and a cosine learning rate decay schedule, with an initial learning rate of 0.002 and a gradual decay over 100 epochs (Loshchilov & Hutter, 2016). I use a batch size of 4096 and a validation split of 20% within the data contained in the four folds used for training. To prevent overfitting of the nuisance functions, I apply early stopping when the MSE validation loss does not improve for 20 epochs and restore the weights that achieve the lowest MSE.

5 Results

5.1 Linear Fixed-Effects Model

Figure 5.1 displays the genre-specific sales rank elasticity estimates from the linear FE model with respect to price. The most price-elastic genres are Philosophy, Psychology, and Cooking. A 1% increase in price worsens their sales rank by approximately 0.51%, 0.25%, and 0.21%, respectively, indicating the largest fall in demand. In contrast, the price-inelastic genres include Fiction: Women, Travel, and Literary Criticism, for which the estimated effects are small and statistically insignificant. Panel A in Table A.1 in the Appendix presents CATEs for all genres, along with their standard errors.

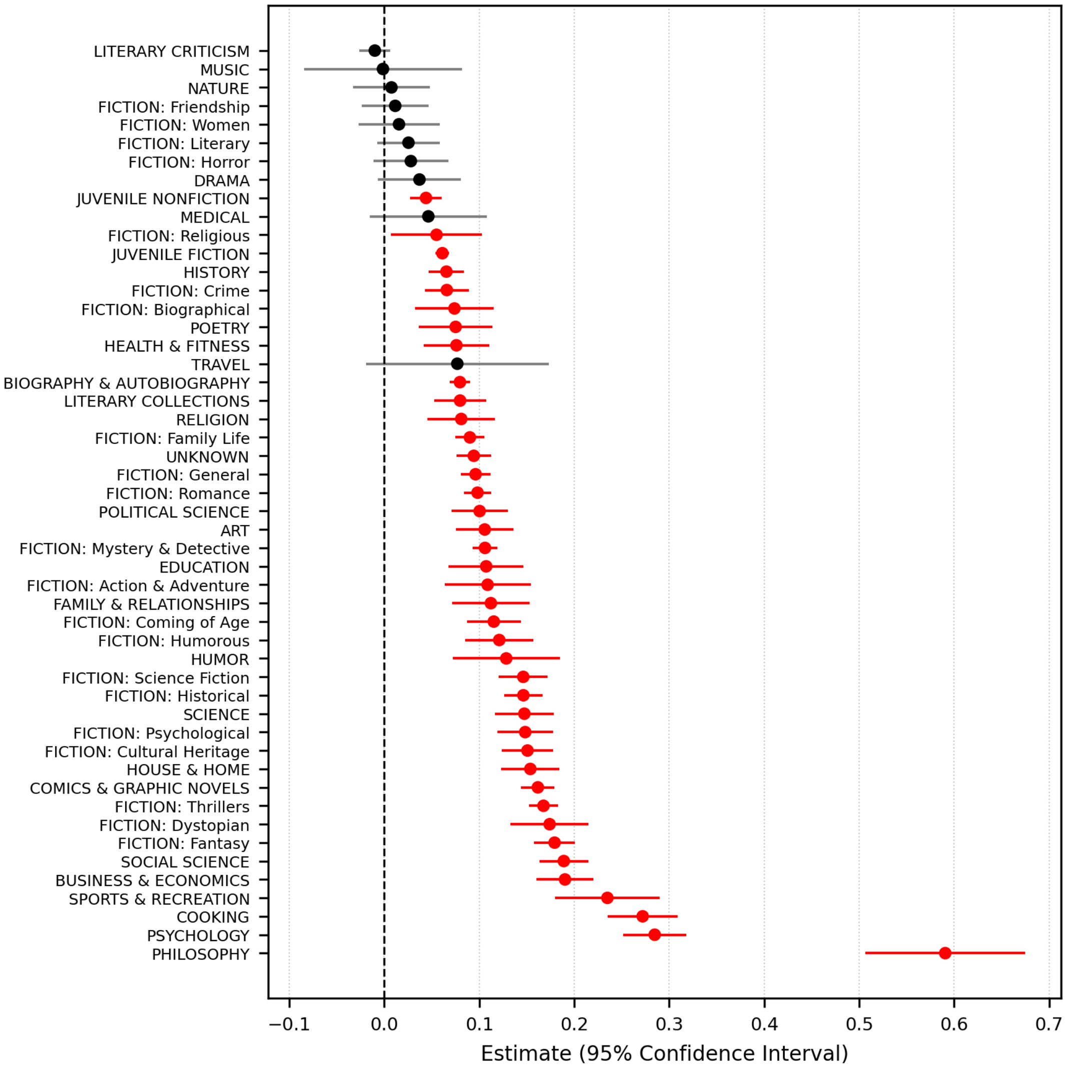

Figure 5.1 Genre-specific sales rank elasticities from the linear FE model with respect to price. Each point represents a CATE estimate for a genre, with horizontal error bars showing 95% confidence intervals. Red color denotes statistically significant estimates at the 5% level, while black color denotes insignificant ones. Positive coefficients correspond to a deterioration in the sales rank. ATE is 0.080 with a standard error of 0.004.

Figure 5.2 displays the genre-specific sales rank elasticity estimates from the linear FE model with respect to star ratings. The most elastic genres are Travel, Cooking, and Health & Fitness. A 1% increase in star ratings improves their sales rank by approximately 0.29%, 0.28%, and 0.28%, respectively, indicating the largest increase in demand. In contrast, the inelastic genres include Philosophy, Religion, Education, and Political Science, for which the estimated effects are small and statistically insignificant. Surprisingly, for Art, the estimated elasticity is positive, meaning that higher star ratings are associated with a deterioration in the sales rank and thus lower demand. Panel A in Table A.2 in the Appendix presents CATEs for all genres, along with their standard errors.

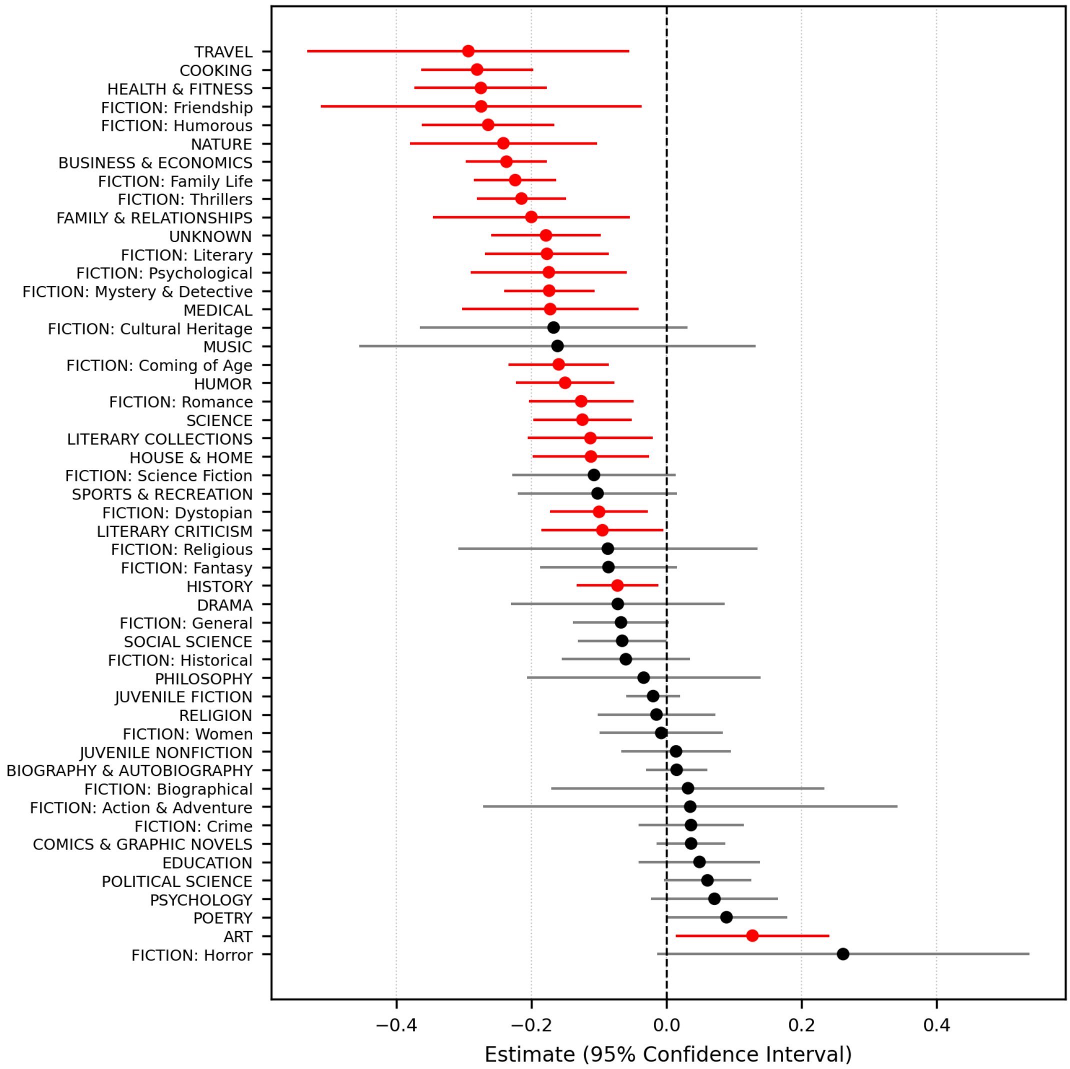

Figure 5.2 Genre-specific sales rank elasticities from the linear FE model with respect to star ratings. Each point represents a CATE estimate for a genre, with horizontal error bars showing 95% confidence intervals. Red color denotes statistically significant estimates at the 5% level, while black color denotes insignificant ones. Negative coefficients correspond to an improvement in the sales rank, while positive coefficients correspond to a deterioration in the sales rank. ATE is -0.082 with a standard error of 0.013.

Figure 5.3 displays the genre-specific sales rank semi-elasticity estimates from the linear FE model with respect to newspaper reviews. In 45 out of 50 genres, sales rank improves significantly following a review by a professional outlet. The most responsive genres are Philosophy, House & Home, and Health & Fitness. Being reviewed by newspapers improves their sales rank by approximately 0.3%, 0.28%, and 0.25%, respectively, during the 20 days following the newspaper review, indicating the largest increase in demand. In contrast, for Juvenile Nonfiction, Fiction: Horror, Music, Drama, and Fiction: Action & Adventure, a newspaper review does not have a statistically significant impact on sales rank. Panel A in Table A.3 in the Appendix presents CATEs for all genres, along with their standard errors.

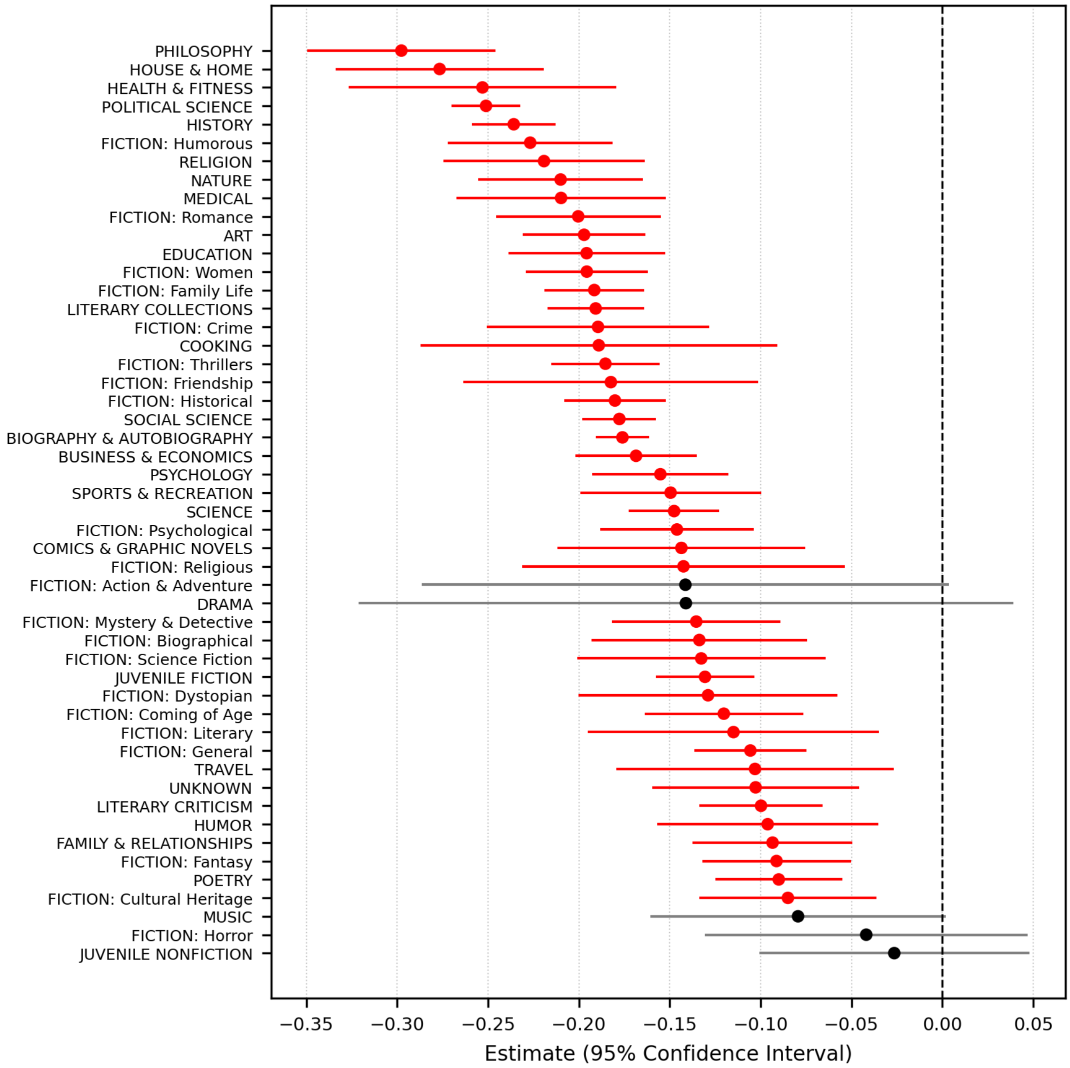

Figure 5.3 Genre-specific sales rank semi-elasticities from the linear FE model with respect to newspaper reviews. Each point represents a CATE estimate for a genre, with horizontal error bars showing 95% confidence intervals. Red color denotes statistically significant estimates at the 5% level, while black color denotes insignificant ones. Negative coefficients correspond to an improvement in the sales rank. ATE is -0.173 with a standard error of 0.005.

5.2 Double/Debiased Machine Learning

Figure 5.4 displays the genre-specific sales rank elasticity estimates from the DML model with respect to price. Similar to the linear FE model, the most price-elastic genres include Philosophy, Psychology, and Cooking. A 1% increase in price worsens their sales rank by approximately 0.59%, 0.28%, and 0.27%, respectively, indicating the largest fall in demand. In contrast, the price-inelastic genres include Literary Criticism, Music, and Nature, for which the estimated effects are small and statistically insignificant. Panel B in Table A.1 in the Appendix presents CATEs for all genres, along with their standard errors.

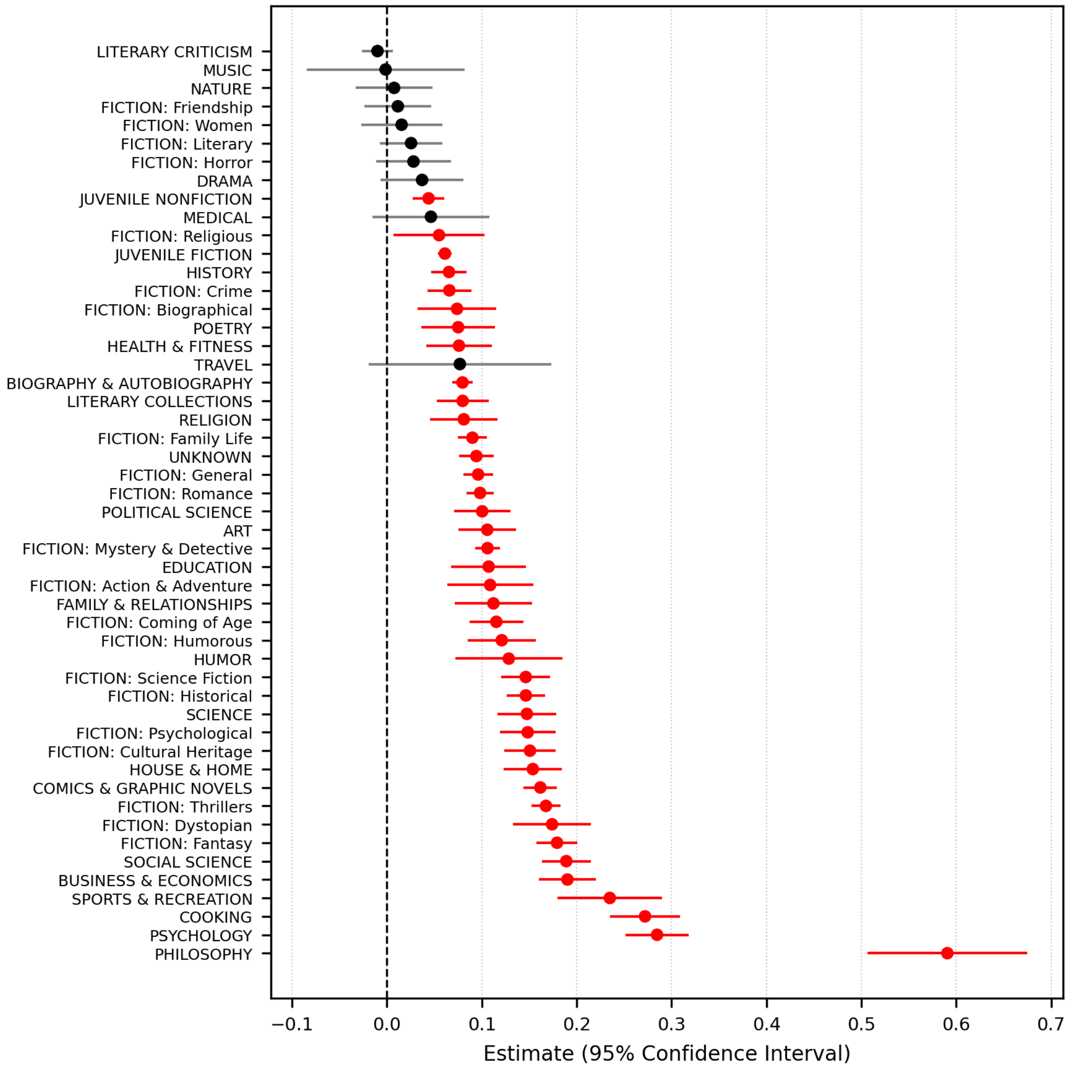

Figure 5.4 Genre-specific sales rank elasticities from the DML model with respect to price. Each point represents a CATE estimate for a genre, with horizontal error bars showing 95% confidence intervals. Red color denotes statistically significant estimates at the 5% level, while black color denotes insignificant ones. Positive coefficients correspond to a deterioration in the sales rank. ATE is 0.096 with a standard error of 0.004.

Figure 5.5 displays the genre-specific sales rank elasticity estimates from the DML model with respect to star ratings. The most elastic genres are Travel, Music, and Health & Fitness. A 1% increase in star ratings improves their sales rank by approximately 0.47%, 0.39%, and 0.3%, respectively, indicating the largest increase in demand. In contrast, the inelastic genres include Education, Political Science, and Philosophy, for which the estimated effects are small and statistically insignificant. Similar to the linear FE model, for Art, the estimated elasticity is again positive, meaning that higher star ratings are associated with a deterioration in the sales rank and thus lower demand. Panel B in Table A.2 in the Appendix presents CATEs for all genres, along with their standard errors.

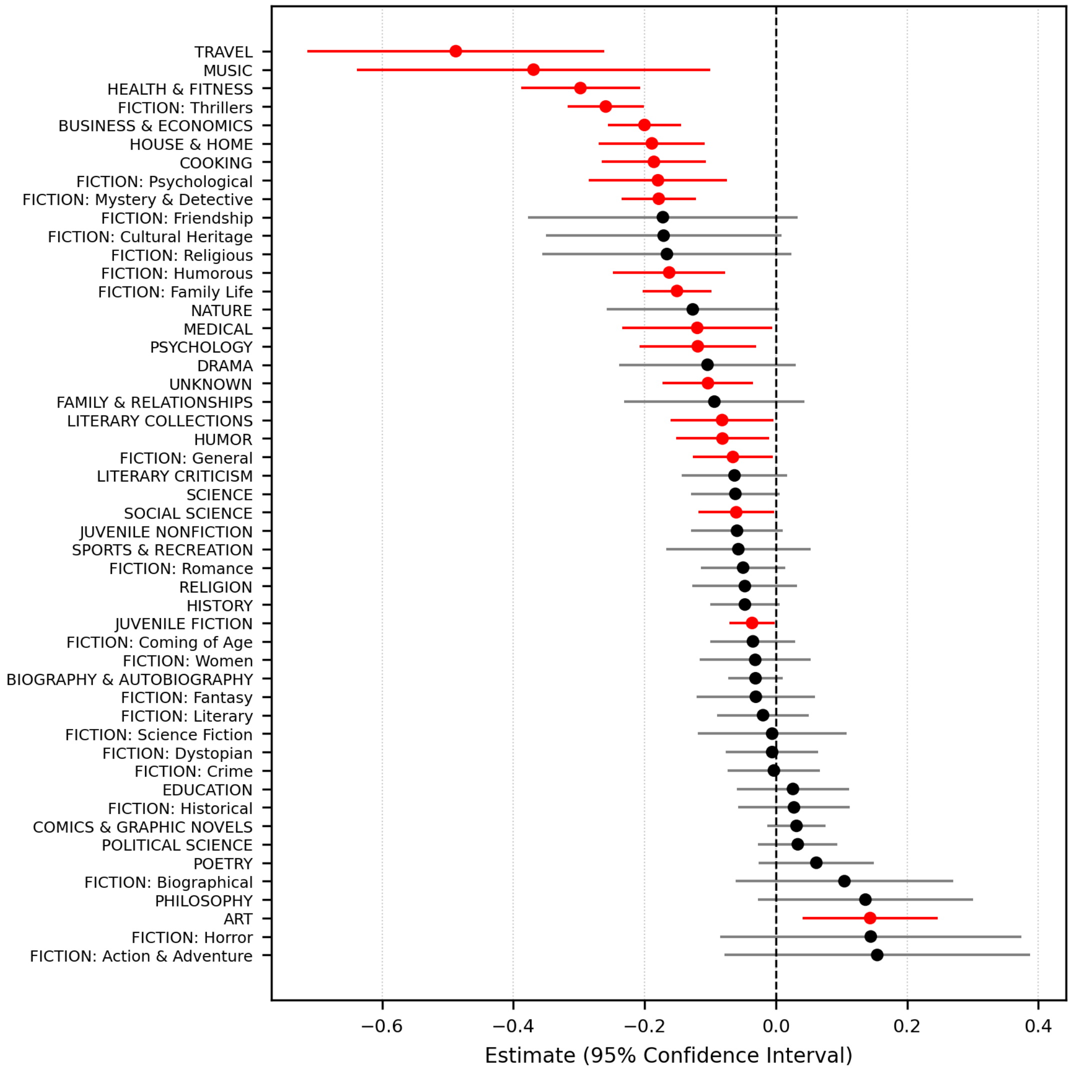

Figure 5.5 Genre-specific sales rank elasticities from the DML model with respect to star ratings. Each point represents a CATE estimate for a genre, with horizontal error bars showing 95% confidence intervals. Red color denotes statistically significant estimates at the 5% level, while black color denotes insignificant ones. Negative coefficients correspond to an improvement in the sales rank, while positive coefficients correspond to a deterioration in the sales rank. ATE is -0.069 with a standard error of 0.01.

Figure 5.6 displays the genre-specific sales rank semi-elasticity estimates from the DML model with respect to newspaper reviews. In 46 out of 50 genres, sales rank improves significantly following a review by a professional outlet. The most responsive genres are Comics & Graphic Novels, Philosophy, and Religion. Being reviewed by newspapers improves their sales rank by approximately 0.55%, 0.27%, and 0.27%, respectively, during the 20 days following the newspaper review, indicating the largest increase in demand. In contrast, for Music, Fiction: Cultural Heritage, Travel, and Cooking, a newspaper review does not have a statistically significant impact on sales rank. Panel B in Table A.3 in the Appendix presents CATEs for all genres, along with their standard errors.

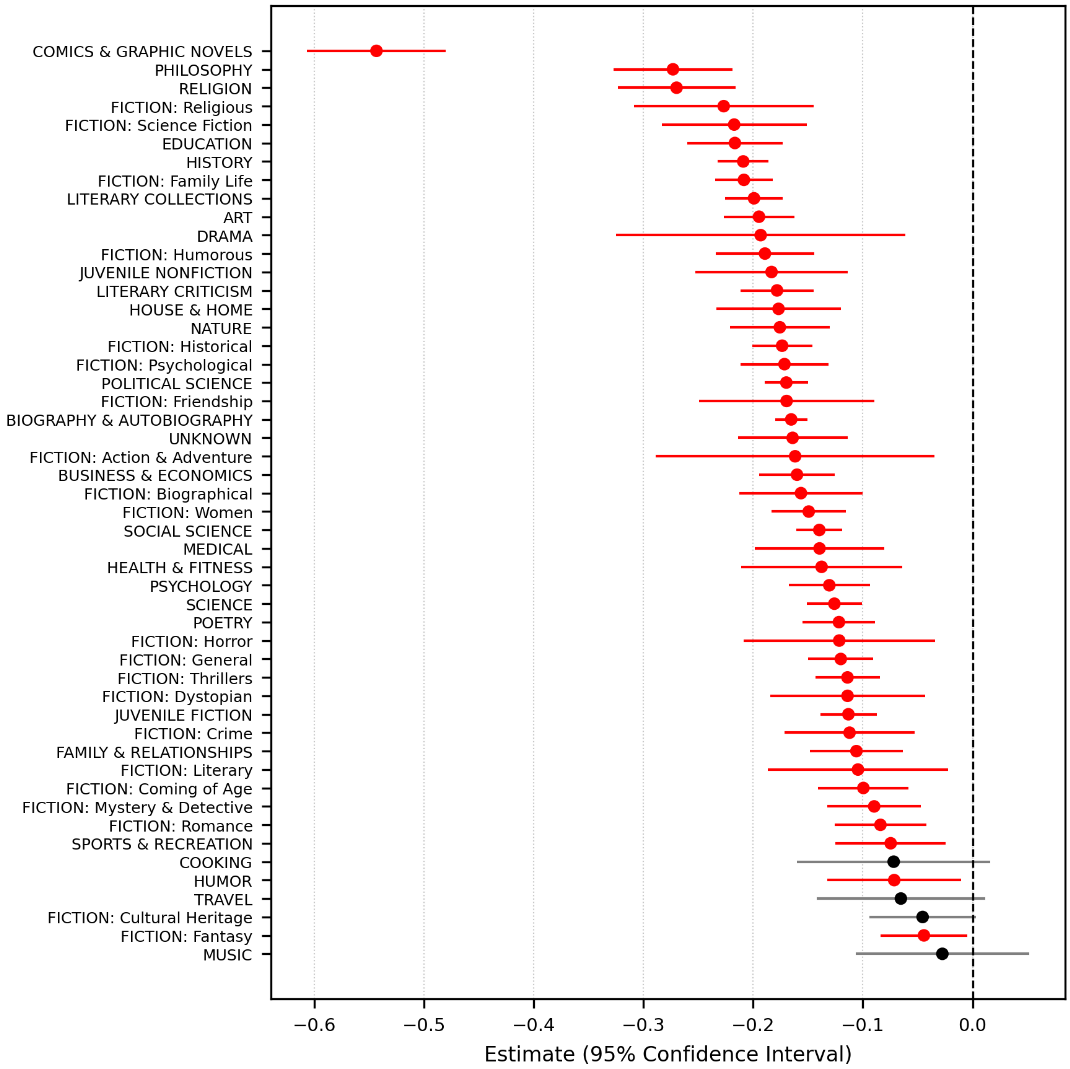

Figure 5.6 Genre-specific sales rank semi-elasticities from the DML model with respect to newspaper reviews. Each point represents a CATE estimate for a genre, with horizontal error bars showing 95% confidence intervals. Red color denotes statistically significant estimates at the 5% level, while black color denotes insignificant ones. Negative coefficients correspond to an improvement in the sales rank. ATE is -0.156 with a standard error of 0.005.

5.3 Comparison of Linear FE and DML Models

In this section, I compare the estimates and standard errors for the linear FE model and the DML model to evaluate whether the DML procedure corrects potential functional form misspecification in the linear FE model. Table 5.1 shows that both the linear FE and DML models deliver similar ATE estimates, suggesting minimal effect of nonlinearities on average effects. Next, I compare the CATE estimates from both models to see whether they yield different results. Significant differences between the linear FE and DML estimates would suggest that genre-specific nonlinear effects are present and may have been averaged out when estimating the ATE. To compare the CATE estimates from the linear FE and DML models, I compute the difference between the two coefficients for each genre g as $\Delta_g = \theta_g^{FE} - \theta_g^{DML}$ and construct confidence intervals using the procedure described in Appendix A.3.

	Panel A: Linear FE		Panel B: DML	
Treatment	Estimate	Std. Error	Estimate	Std. Error
Price	0.080	0.004	0.096	0.004
Star Ratings	-0.082	0.013	-0.069	0.010
Newspaper Reviews	-0.173	0.005	-0.156	0.005

Table 5.1 ATE estimates of the effects of price, star ratings, and newspaper reviews on sales rank from the linear FE and DML models. Standard errors are clustered by edition-platform.

Figure 5.7 shows the differences between linear FE and DML genre-specific price elasticity estimates. For 12 out of the 50 genres, the difference is statistically significant, suggesting that some CATE estimates from the linear FE model may have been biased due to previously unmodeled nonlinear relationships. Comparing the standard errors of the CATE estimates from the linear FE model with the corresponding standard errors from the DML model for each genre reveals that DML produced lower standard errors in 47 out of 50 genres. This result points to a systematic increase in efficiency after accounting for the nonlinear effects of confounders.

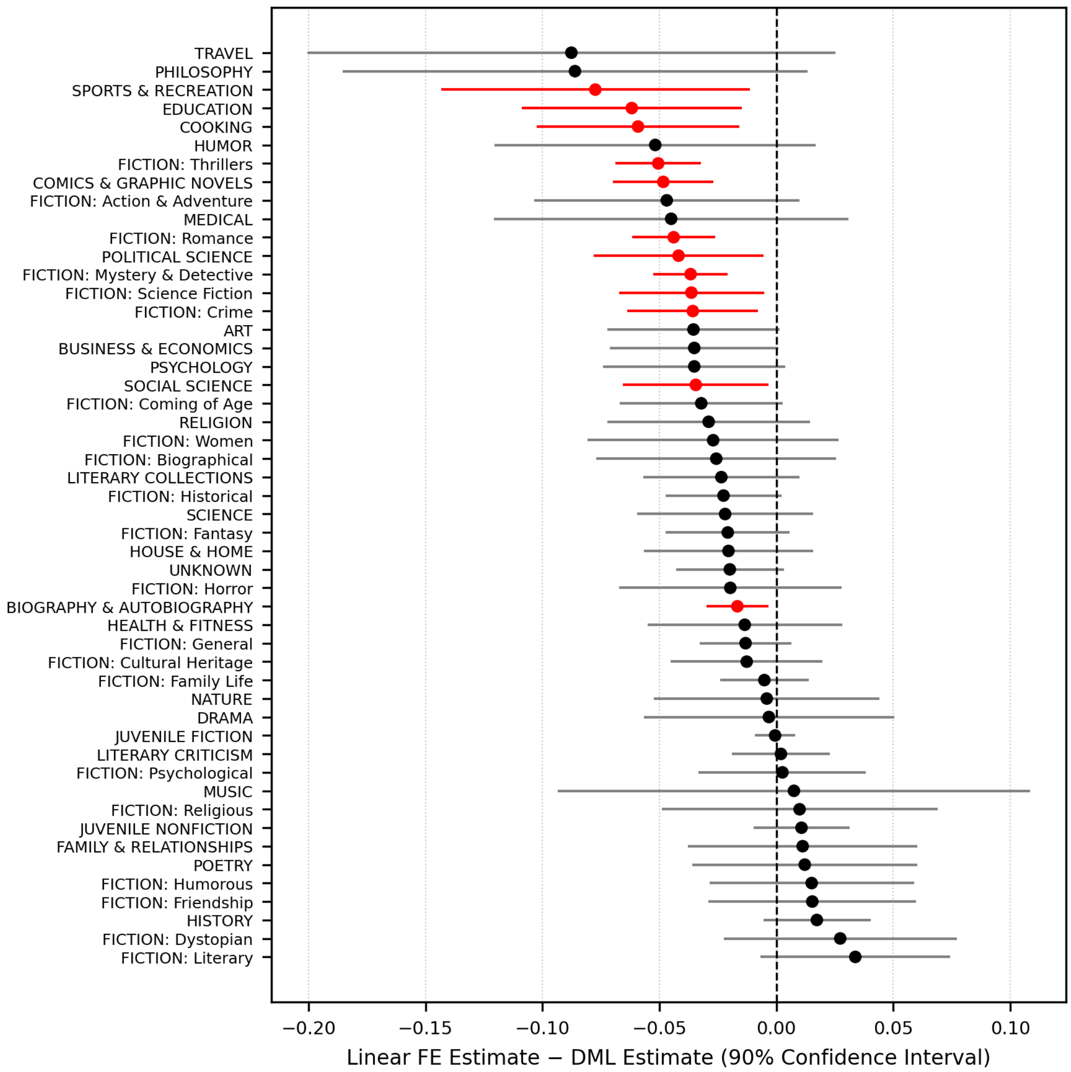

Figure 5.7 Difference between linear FE and DML genre-specific price elasticity estimates. Horizontal error bars show 90% confidence intervals. Confidence intervals assume zero covariance between FE and DML estimates. Red color denotes statistically significant differences at the 10% level, while black denotes insignificant ones.

Figure 5.8 shows the differences between linear FE and DML genre-specific elasticity estimates for star ratings. For 5 out of the 50 genres, the difference is statistically significant, suggesting that only a few CATE estimates from the linear FE model may have been biased due to previously unmodeled nonlinear relationships. Comparing the standard errors of the CATE estimates from the linear FE model with the corresponding standard errors from the DML model for each genre reveals that DML produced lower standard errors across all genres. This result indicates an increase in efficiency across the board after accounting for the nonlinear effects of confounders.

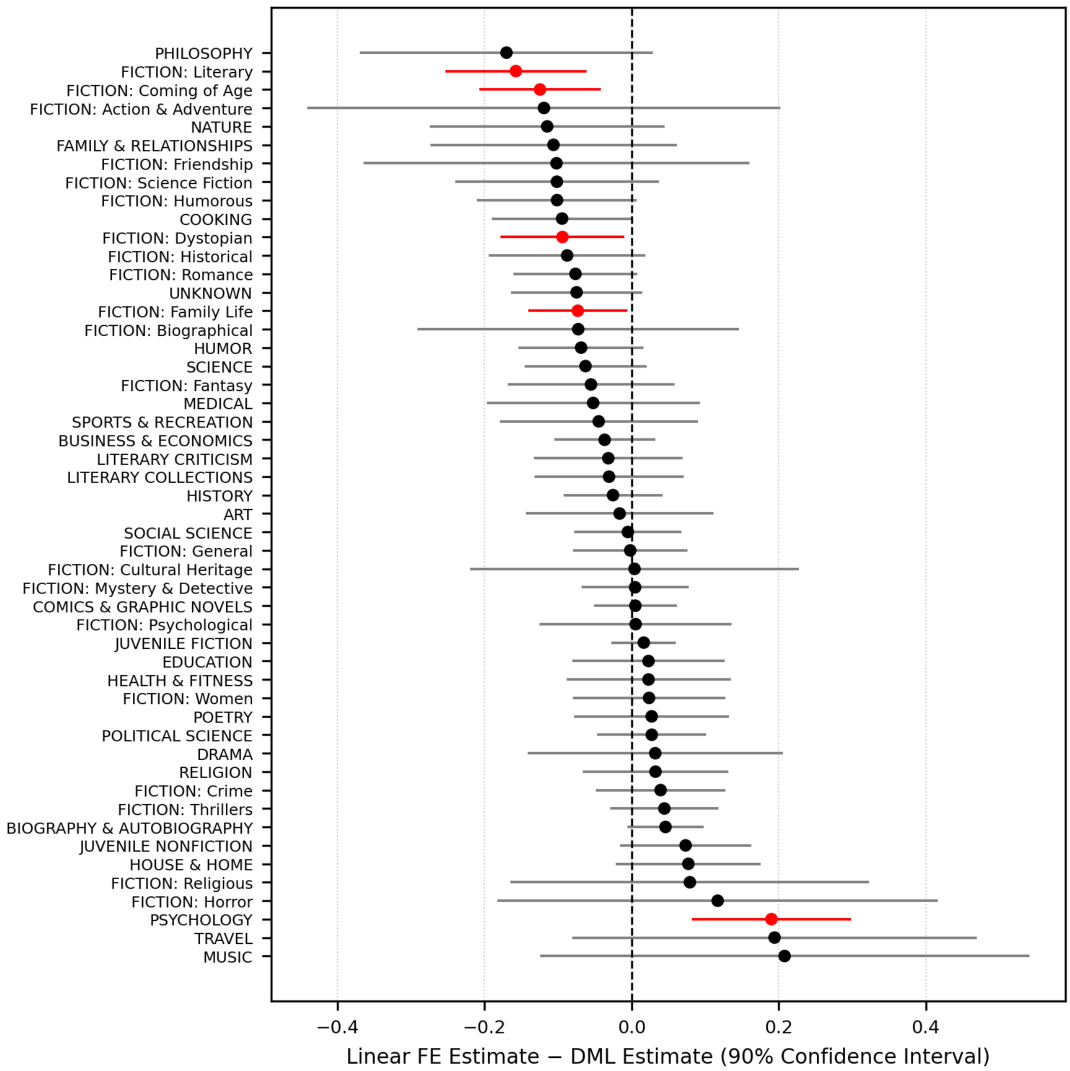

Figure 5.8 Difference between linear FE and DML genre-specific elasticity estimates for star ratings. Horizontal error bars show 90% confidence intervals. Confidence intervals assume zero covariance between FE and DML estimates. Red color denotes statistically significant differences at the 10% level, while black denotes insignificant ones.

Figure 5.9 shows the differences between linear FE and DML genre-specific semi-elasticity estimates for newspaper reviews. For 15 out of the 50 genres, the difference is statistically significant, suggesting that almost one-third of CATE estimates from the linear FE model may have been biased due to previously unmodeled nonlinear relationships. Comparing the standard errors of the CATE estimates from the linear FE model with the corresponding standard errors from the DML model for each genre reveals that DML produced lower standard errors in 35 out of 50 genres. This result points to an increase in efficiency after accounting for the nonlinear effects of confounders.

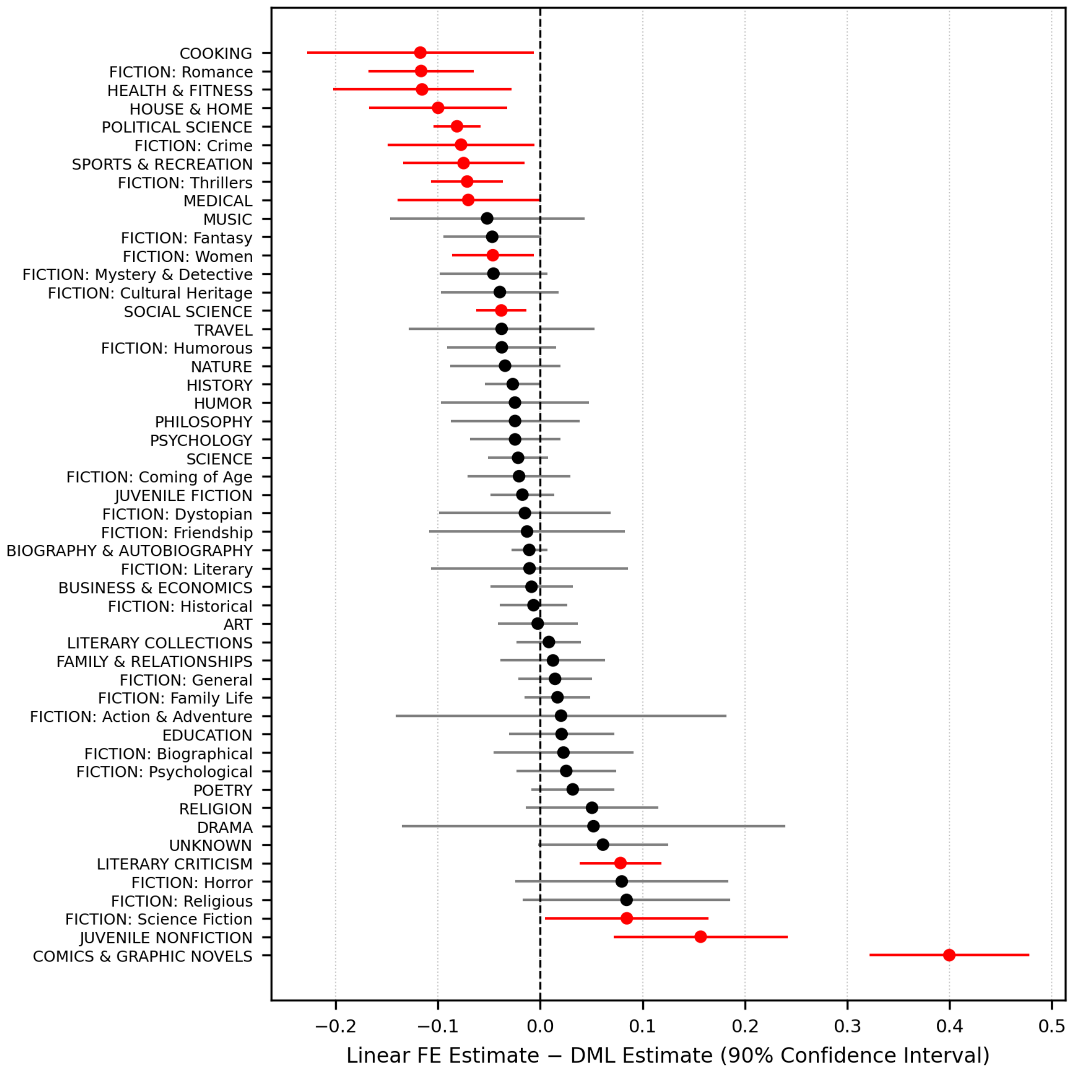

Figure 5.9 Difference between linear FE and DML genre-specific semi-elasticity estimates for newspaper reviews. Horizontal error bars show 90% confidence intervals. Confidence intervals assume zero covariance between FE and DML estimates. Red color denotes statistically significant differences at the 10% level, while black denotes insignificant ones.

5.4 Visualizing Nonlinear Confounding

In this section, I visualize the relationships between the outcome $\widetilde{\ln}(rank_{jct})$ and each of the confounding variables, namely, $\widetilde{\ln}(rank_{jc,t-1})$, $\widetilde{\ln}(reviews_{jct})$, $\widetilde{U}_{jt}$, and $\widetilde{S}_{jt}$ to assess whether differences in CATE estimates between the linear FE and DML models are driven by nonlinear relationships in these confounders. To this end, I use accumulated local effects (ALE) plots that illustrate how a specific feature impacts the average prediction of an ML model while keeping all other features at their observed values. ALE plots are a faster and unbiased alternative to partial dependence plots, which can become biased if features are correlated (Molnar, 2025). To draw ALE plots, I set up a neural network model $g(\mathbf{X})$ with $\mathbf{X} = (\widetilde{\ln}(p_{jct}), \widetilde{\ln}(stars_{jct}), \widetilde{news}_{jt}, \widetilde{\ln}(rank_{jc,t-1}), \widetilde{\ln}(reviews_{jct}), \widetilde{U}_{jt}, \widetilde{S}_{jt})$ to predict the outcome $\widetilde{\ln}(rank_{jct})$. The neural network uses the same architecture as I described in subsection 4.2.1. This prediction problem is similar to the outcome equation I estimate in the DML framework; however, here, all variables, including the treatments, enter the neural network in a fully flexible way. Because a neural network can approximate any measurable function to any desired degree of accuracy, a nonlinear ALE curve for a confounding variable provides direct evidence of a nonlinear effect of that confounder on the outcome (Hornik, Stinchcombe, & White, 1989; Molnar, 2025). I describe the math behind the computation of ALE plots in Appendix A.4.

Figure 5.10 shows the ALE for the effect of $\widetilde{\ln}(rank_{jc,t-1})$ on the neural network prediction of the outcome. The curve is approximately linear, suggesting little to no bias from functional form misspecification when modeling the impact of $\widetilde{\ln}(rank_{jc,t-1})$ linearly.

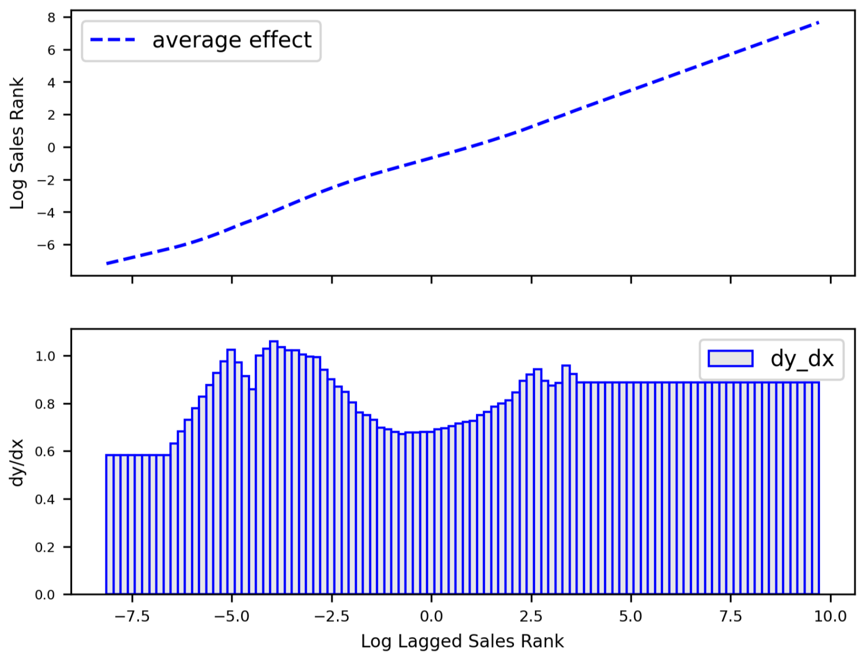

Figure 5.10 ALE of $\widetilde{\ln}(rank_{jc,t-1})$ on the neural network prediction of $\widetilde{\ln}(rank_{jct})$

Figure 5.11 shows the ALE for the effect of $\widetilde{\ln}(reviews_{jct})$ on the neural network prediction of the outcome. The curve is highly nonlinear, indicating a substantial bias from misspecifying the functional form if the impact of $\widetilde{\ln}(reviews_{jct})$ is modeled linearly. It also has a clear interpretation, showing that sales rank gradually improves as the number of reviews increases up to a certain point, after which it worsens exponentially. This pattern indicates that book editions with very high numbers of reviews are typically those where the reviews are mostly negative, which translates into lower demand.

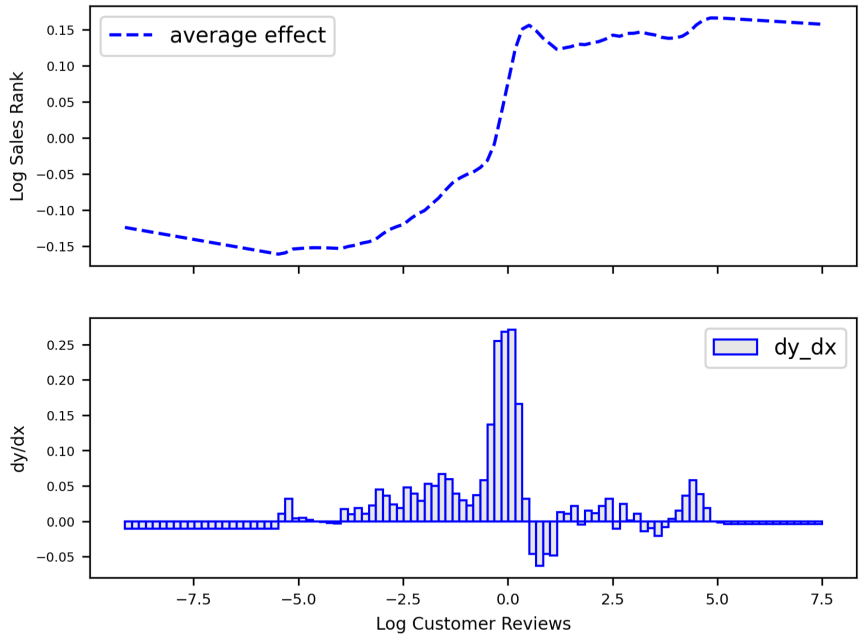

Figure 5.11 ALE of $\widetilde{\ln}(reviews_{jct})$ on the neural network prediction of $\widetilde{\ln}(rank_{jct})$

Figure 5.12 displays the ALE for the effect of $\widetilde{U}_{jt}$ on the neural network prediction of the outcome. The curve is highly nonlinear, indicating a substantial bias from misspecifying the functional form when the impact of $\widetilde{U}_{jt}$ is modeled linearly.

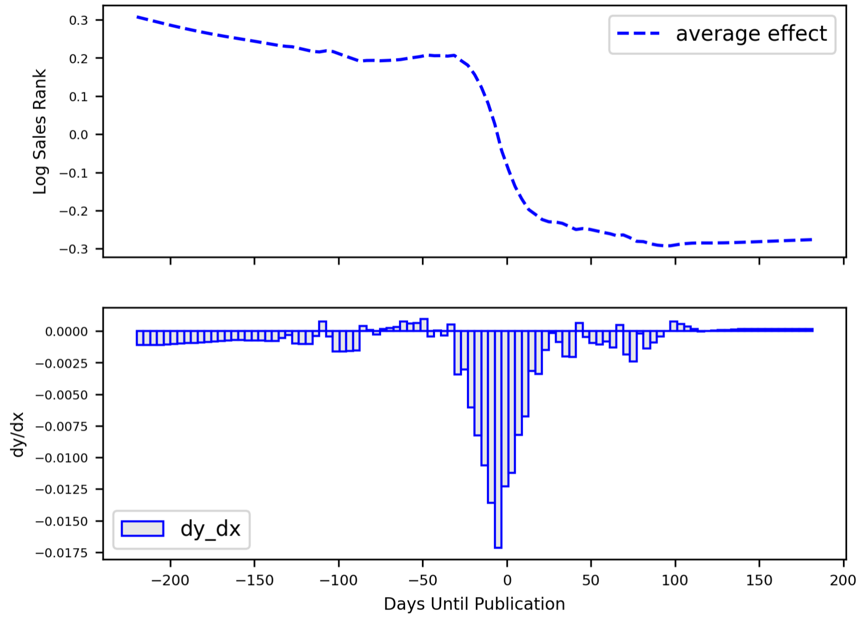

Figure 5.12 ALE of $\widetilde{U}_{jt}$ on the neural network prediction of $\widetilde{\ln}(rank_{jct})$

Figure 5.13 shows the ALE for the effect of $\widetilde{S}_{jt}$ on the neural network prediction of the outcome. The curve is highly nonlinear, following a sinusoidal pattern, which indicates a severe bias from misspecifying the functional form when the impact of $\widetilde{S}_{jt}$ is modeled linearly.

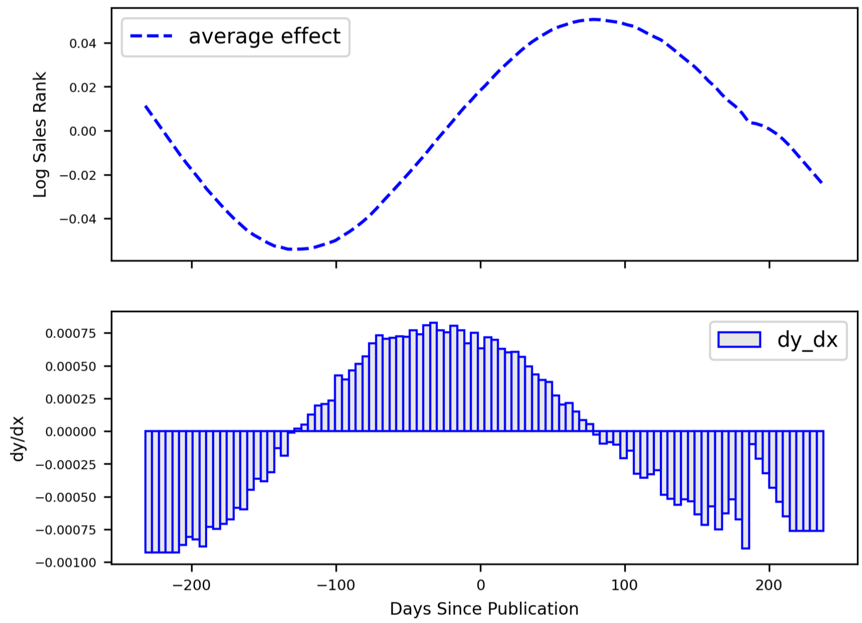

Figure 5.13 ALE of $\widetilde{S}_{jt}$ on the neural network prediction of $\widetilde{\ln}(rank_{jct})$

Discussion

In this chapter, I summarize and interpret the results of my thesis. I also discuss its limitations. I interpret the sales rank CATE elasticities using the DML model, which constitutes the main specification of this thesis because it accounts for the nonlinear effects of confounding variables. Overall, my estimates using either the linear FE or the DML model indicate substantial heterogeneity across genres with respect to prices, star ratings, and newspaper reviews, as genre-specific CATEs differ significantly from ATEs. This heterogeneity suggests that treating the book market as a whole can be inefficient, and that segmentation strategies tailored to genre-specific demand patterns are more appropriate.

Genre-specific CATEs for price using the DML model show that the price-elastic genres include Philosophy, Psychology, Cooking, Business & Economics, Sports & Recreation, and Social Science. In these categories, readers tend to shop by topic rather than by a specific title or author, so a price increase for one book leads them to substitute it for another similar title. Additionally, Cooking, Business & Economics, and Sports & Recreation have many free options available online, such as YouTube videos, making it easier for consumers to switch away if prices rise. In these genres, even slight price reductions are likely to yield noticeable improvements in the sales rank, reflecting the strong demand responsiveness to price changes. The price-inelastic genres include Literary Criticism, Music, Nature, Medical, Drama, Fiction: Friendship, Fiction: Literary, Fiction: Horror, and Fiction: Women. In these genres, readers tend to care more about a particular narrative or author, leading to little substitution when the price changes. As a result, modest price increases are unlikely to significantly impact the sales rank, suggesting that demand responds only weakly to pricing changes.

Genre-specific CATEs for star ratings using the DML model show that the genres whose sales rank responds the most to changes in star ratings include Travel, Music, Health & Fitness, Fiction: Thrillers, Business & Economics, House & Home, and Cooking. These categories often attract readers who rely on star ratings as an important signal of expected quality, particularly when they do not have strong preferences for a specific author or title. Marketers promoting books in these genres may benefit from emphasizing high star ratings, as demand is relatively responsive to changes in star ratings. In contrast, genres whose sales rank shows little response to changes in star ratings include Fiction: Action & Adventure, Fiction: Horror, Philosophy, Fiction: Biographical, Poetry, and Political Science. Readers in these genres tend to have their own preferences for particular authors or styles, so star ratings play a limited role in their purchase decisions. Highlighting high star ratings in marketing campaigns for these genres is therefore unlikely to produce meaningful changes in demand.

Genre-specific CATEs for newspaper reviews using the DML model show that the genres whose sales rank responds the most to a professional review include Comics & Graphic Novels, Philosophy, Religion, Fiction: Religious, Fiction: Science Fiction, Education, and History. These categories often attract more specialized or highly engaged readers, for whom expert judgment offers valuable guidance when choosing among titles. In these genres, highlighting that a book has been reviewed by a professional newspaper, for example, by placing review excerpts on the book cover, can significantly boost demand. In contrast, genres whose sales rank responds only slightly to professional reviews include Music, Fiction: Cultural Heritage, Travel, and Cooking. These genres generally appeal to broader audiences who rely primarily on peer evaluations and practical feedback. In these categories, consumer reviews provide the most relevant information for assessing usefulness or enjoyment, while professional reviews have only a limited impact. An interesting pattern supports this interpretation: Music, Travel, and Cooking, whose sales rank does not respond significantly to professional reviews, were among the most responsive genres to changes in star ratings. For these genres, emphasizing professional reviews on the book cover is unlikely to significantly impact demand, as star ratings are more influential in purchasing decisions.

To evaluate whether the DML model produces estimates significantly different from those of the linear FE model, I compared the genre-specific CATEs across both specifications. For all three treatments, the estimates from the DML and linear FE models differ significantly for a substantial number of genres, indicating that some genre-specific CATEs may have been distorted by unmodeled nonlinear relationships in the linear FE model that the DML approach can capture. I visualized the nonlinear confounding structure using the ALE plots, which show that the number of customer ratings, the number of days until the publication, and the number of days since publication have a nonlinear effect on sales rank. Additionally, for all treatments, the vast majority of genre-specific standard errors are lower under the DML model than under the linear FE model, pointing to a systematic increase in efficiency after accounting for the nonlinear effects of confounders. By accounting for the nonlinear structure in the control variables, the residual variation decreases, which in turn reduces the variance of the estimator (Strittmatter, 2025).

Estimating genre-specific CATEs incorrectly due to unmodeled nonlinear relationships in the linear FE model can lead to a misallocation of marketing resources toward campaigns that are ineffective within the targeted genres. Several comparisons between linear FE and DML estimates illustrate this point. For Fiction: Literary, the sales-rank elasticity with respect to star ratings is -0.18 and statistically significant in the linear FE model, but only -0.02 and statistically insignificant in the DML model. Similarly, for Fiction: Coming of Age, the linear FE estimate is -0.16 and significant, whereas the DML estimate is -0.04 and insignificant at the 5% level. Based on these estimates, the linear FE model would encourage marketers in these genres to emphasize star ratings, while the DML results suggest that such efforts are unlikely to shift demand in a meaningful way once nonlinear confounding is accounted for. This pattern is consistent with my interpretation that readers in more niche genres often have strong preferences for particular authors or styles, so customer ratings play a limited role in purchase decisions. In contrast, these more specialized readers tend to be more receptive to professional reviews, as reflected in newspaper review semi-elasticities of -0.12 for Fiction: Literary and -0.12 for Fiction: Coming of Age. Likewise, the DML estimates point to a different optimal marketing strategy for Cooking than the linear FE model. The linear FE model yields a sizable and statistically significant effect for newspaper reviews, with a semi-elasticity of -0.19, while the DML estimate is much smaller, at -0.07, and statistically insignificant at the 5% level. In this case, the linear FE specification would motivate campaigns that feature professional reviews, whereas the DML results indicate that such reviews matter little for demand in this genre. At the same time, Cooking is among the most responsive genres to star ratings in the DML model, with a sales-rank elasticity of -0.19, suggesting that marketing resources are more effectively allocated toward strategies that highlight customer ratings rather than newspaper coverage.

A surprising result emerged when estimating the impact of star ratings on sales rank for the Art genre. The sales rank elasticity was consistently positive when using either of the linear FE or DML models, suggesting that higher star ratings are associated with a deterioration in sales rank and, thus, lower demand. This economically counterintuitive result may require further examination with a larger sample of art books and richer covariates to confirm whether the relationship between star ratings and demand is indeed negative or whether it is a result of omitted factors specific to the Art genre.

In my analysis, I use sales rank as a proxy for consumer demand because the relationship between the natural logarithm of sales rank and the natural logarithm of demand volume is approximately linear (Chevalier & Goolsbee, 2003). This means that the estimated sales rank elasticities with respect to price, star ratings, and newspaper reviews differ from the true demand elasticities only by a scalar transformation (Reimers & Waldfogel, 2021). While sales rank elasticities allow me to identify which genres are more responsive to changes in the treatment variables, they provide only ordinal information about relative demand responsiveness rather than cardinal information about the magnitude of quantity changes. Further research could provide marketers with more detailed guidance on pricing strategies by analyzing actual sales volumes, which would make it possible to measure demand responses in cardinal rather than merely ordinal terms. Access to such data would allow researchers to quantify the magnitude of demand changes, complementing the relative responsiveness captured by sales rank elasticities.

Additionally, although the DML framework accounts for covariates flexibly, it still imposes a linear specification on the treatment variables. If the true relationship between treatments and the outcome is nonlinear, the estimated coefficients may be biased due to omitted nonlinear treatment terms. A fully non-parametric approach, such as a generalized random forest, could, in principle, accommodate such nonlinearities in the treatment effect (Athey, Tibshirani, & Wager, 2019). However, addressing nonlinearities in this way comes at a cost, because a generalized random forest does not deliver the structural genre-specific elasticities that are the focus of my analysis. When the treatment effect is nonlinear, it naturally varies with the covariates, so the forest delivers a different local effect for each observation rather than a single parameter per genre. Even if one averaged the observation-level CATEs within a genre, the resulting averages would not correspond to the structural genre-specific elasticities that I aim to estimate, because the forest does not constrain treatment effects to be constant within genres. For this reason, tree-based estimators cannot provide the interpretable, genre-level elasticities required for my analysis, which motivates the use of DML with treatment-genre interactions instead.

Conclusion

In this thesis, I examine how demand elasticities with respect to price, Amazon star ratings, and newspaper reviews vary across book genres in the Amazon book marketplace. Following the academic literature, I use Amazon sales rank as a proxy for consumer demand, where a lower rank indicates higher demand and a higher rank indicates lower demand. I use a high-frequency panel with daily data on book demand in 2018 from the Amazon domains in the United States, Canada, and the United Kingdom. I estimate two types of models: a linear fixed-effects (FE) model, which serves as a benchmark, and a double/debiased machine learning (DML) model, which is the main specification because it flexibly accounts for nonlinear effects of control variables while preserving the interpretability of treatment effects as elasticities.

My core finding is that consumer demand in the Amazon book marketplace is heterogeneous, as sales rank responsiveness to changes in price, star ratings, and professional newspaper reviews differs substantially across genres. This heterogeneity suggests that treating the book market as a whole can be inefficient, and that segmentation strategies tailored to genre-specific demand patterns are more appropriate. Building on this insight, my analysis yields several practical implications for booksellers and marketers. In particular, I identify genres that are relatively price-inelastic, in which booksellers can raise prices with only limited adverse effects on demand. I further point to genres in which demand is especially responsive to changes in star ratings, indicating that marketing strategies emphasizing high customer ratings are particularly effective in these categories. Finally, I highlight genres in which professional newspaper reviews have the strongest influence on demand, suggesting that featuring such reviews, for example, through excerpts on book covers, can substantially increase demand in these segments.

Beyond its marketing implications, this thesis offers a methodological insight for estimating heterogeneous demand elasticities in the presence of nonlinear confounding. Genre-specific elasticity estimates for price, star ratings, and newspaper reviews obtained from the linear FE model differ significantly for a substantial number of genres relative to the corresponding estimates in the DML model. This divergence indicates that some genre-specific elasticity estimates may be distorted when nonlinear relationships are not adequately captured in the linear FE model but are accounted for in the DML approach. In several cases, the linear FE and DML models produced estimates that differ enough to imply different optimal marketing strategies across genres. Moreover, the DML specification typically delivers more precise genre-specific estimates than the linear FE model, consistent with efficiency gains from accounting for nonlinear confounding. Taken together, these findings motivate using DML as the main empirical specification of the thesis and interpreting its genre-specific elasticity estimates as the primary results.

One important limitation of this thesis is that I measure demand using Amazon sales rank rather than observed quantities sold. While sales rank is a standard proxy in the literature and is approximately log-linearly related to demand volume (Chevalier & Goolsbee, 2003), the estimated elasticities primarily provide ordinal information about relative demand responsiveness across genres. They indicate which genres are more or less responsive to changes in price, star ratings, and newspaper reviews, but they do not directly quantify the cardinal magnitude of changes in quantities sold. Future research with access to sales volumes could measure demand responses directly, allowing researchers to express elasticity estimates in terms of actual quantities sold and to provide marketers and booksellers with clearer guidance when designing genre-based segmentation strategies.

Bibliography

Abrevaya, J., Hsu, Y.-C., & Lieli, R. P. (2015). Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, *33*(4), 485–505. https://doi.org/10.1080/07350015.2014.975555

Allenby, G. M., & Rossi, P. E. (1998). Marketing models of consumer heterogeneity. *Journal of Econometrics*, *89*(1), 57–78. https://doi.org/10.1016/S0304-4076(98)00055-4

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2). https://doi.org/10.1214/18-AOS1709

Bach, P., Kurz, M. S., Chernozhukov, V., Spindler, M., & Klaassen, S. (2024). DoubleML: An object-oriented implementation of double machine learning in r. *Journal of Statistical Software*, *108*(3). https://doi.org/10.18637/jss.v108.i03

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, *81*(2), 608–650. https://doi.org/10.1093/restud/rdt044

Belloni, A., Chernozhukov, V., Hansen, C., & Kozbur, D. (2016). Inference in high-dimensional panel models with an application to gun control. *Journal of Business & Economic Statistics*, *34*(4), 590–605. https://doi.org/10.1080/07350015.2015.1102733

Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, *25*(2), 242. https://doi.org/10.2307/2555829

Blackwell, M., & Olson, M. P. (2022). Reducing model misspecification and bias in the estimation of interactions. *Political Analysis*, *30*(4), 495–514. https://doi.org/10.1017/pan.2021.19

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees* (1st ed.). Routledge. https://doi.org/10.1201/9781315139470

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), 1–68. https://doi.org/10.1111/ectj.12097

Chevalier, J., & Goolsbee, A. (2003). Measuring prices and price competition online: Amazon.com and BarnesandNoble.com. *Quantitative Marketing and Economics*, *1*(2), 203–222. https://doi.org/10.1023/A:1024634613982

Chevalier, J., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, *43*(3), 345–354. https://doi.org/10.1509/jmkr.43.3.345

Fuhr, J., & Papies, D. (2024). Double machine learning meets panel data - promises, pitfalls, and potential solutions. *arXiv*. https://doi.org/10.48550/ARXIV.2409.01266

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366. https://doi.org/10.1016/0893-6080(89)90020-8

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv*. https://doi.org/10.48550/ARXIV.1412.6980

Loshchilov, I., & Hutter, F. (2016). SGDR: Stochastic gradient descent with warm restarts. *arXiv*. https://doi.org/10.48550/ARXIV.1608.03983

Molnar, C. (2025). *Interpretable machine learning: A guide for making black box models explainable* (3rd ed.). Christoph Molnar.

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87

Reimers, I., & Waldfogel, J. (2021). Digitization and pre-purchase information: The causal and welfare impacts of reviews and crowd ratings. *American Economic Review*, *111*(6), 1944–1971. https://doi.org/10.1257/aer.20200153

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, *56*(4), 931–954. https://doi.org/10.2307/1912705

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Strittmatter, A. (2025). Machine learning for causal inference in economics. *IZA World of Labor*. https://doi.org/10.15185/izawol.516

Swamy, P. A. V. B. (1970). Efficient inference in a random coefficient regression model. *Econometrica*, *38*(2), 311. https://doi.org/10.2307/1913012

Szandała, T. (2021). Review and comparison of commonly used activation functions for deep neural networks. In *Bio-inspired neurocomputing* (1st ed., pp. 203–224). Springer Singapore. https://doi.org/10.1007/978-981-15-5495-7_11

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28. https://doi.org/10.1257/jep.28.2.3

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839

Wooldridge, J. M. (2020). *Introductory econometrics: A modern approach* (7th ed.). Cengage Learning.

A Appendix

A.1 Tables

Genre	Panel A: Linear FE Estimate	Panel A: Linear FE Std. Error	Panel B: DML Estimate	Panel B: DML Std. Error
ART	0.070	0.016	0.106	0.016
BIOGRAPHY & AUTOBIOGRAPHY	0.063	0.006	0.080	0.006
BUSINESS & ECONOMICS	0.155	0.016	0.190	0.015
COMICS & GRAPHIC NOVELS	0.113	0.010	0.162	0.009
COOKING	0.213	0.019	0.272	0.019
DRAMA	0.034	0.024	0.037	0.022
EDUCATION	0.045	0.021	0.107	0.020
FAMILY & RELATIONSHIPS	0.123	0.022	0.112	0.021
FICTION: Action & Adventure	0.062	0.026	0.109	0.023
FICTION: Biographical	0.048	0.023	0.074	0.021
FICTION: Coming of Age	0.083	0.016	0.115	0.014
FICTION: Crime	0.030	0.012	0.066	0.012
FICTION: Cultural Heritage	0.138	0.014	0.151	0.014
FICTION: Dystopian	0.201	0.022	0.174	0.021
FICTION: Family Life	0.085	0.008	0.090	0.008
FICTION: Fantasy	0.158	0.012	0.179	0.011
FICTION: Friendship	0.027	0.020	0.011	0.018
FICTION: General	0.083	0.009	0.096	0.008
FICTION: Historical	0.124	0.011	0.146	0.010
FICTION: Horror	0.008	0.021	0.028	0.020
FICTION: Humorous	0.136	0.019	0.121	0.018
FICTION: Literary	0.059	0.018	0.025	0.017
FICTION: Mystery & Detective	0.069	0.007	0.106	0.007
FICTION: Psychological	0.151	0.016	0.148	0.015
FICTION: Religious	0.065	0.026	0.055	0.024
FICTION: Romance	0.054	0.008	0.098	0.007
FICTION: Science Fiction	0.110	0.014	0.146	0.013
FICTION: Thrillers	0.117	0.008	0.168	0.008
FICTION: Women	-0.012	0.025	0.015	0.022
HEALTH & FITNESS	0.062	0.018	0.076	0.018
HISTORY	0.083	0.010	0.065	0.010
HOUSE & HOME	0.133	0.016	0.154	0.016
HUMOR	0.077	0.030	0.128	0.029
JUVENILE FICTION	0.061	0.004	0.061	0.004
JUVENILE NONFICTION	0.055	0.009	0.044	0.009
LITERARY COLLECTIONS	0.056	0.015	0.080	0.014
LITERARY CRITICISM	-0.008	0.010	-0.010	0.008
MEDICAL	0.001	0.034	0.046	0.032
MUSIC	0.006	0.045	-0.001	0.042
NATURE	0.003	0.021	0.008	0.021
PHILOSOPHY	0.505	0.043	0.591	0.043
POETRY	0.087	0.022	0.075	0.020
POLITICAL SCIENCE	0.059	0.016	0.100	0.015
PSYCHOLOGY	0.250	0.017	0.285	0.017
RELIGION	0.052	0.019	0.081	0.018
SCIENCE	0.126	0.017	0.147	0.016
SOCIAL SCIENCE	0.155	0.014	0.189	0.013
SPORTS & RECREATION	0.158	0.029	0.235	0.028
TRAVEL	-0.011	0.049	0.077	0.049
UNKNOWN	0.074	0.011	0.094	0.009
ATE	0.080	0.004	0.096	0.004

Table A.1 Genre-specific sales rank elasticities with respect to price. Standard errors are clustered by edition-platform.

Genre	Panel A: Linear FE		Panel B: DML	
	Estimate	Std. Error	Estimate	Std. Error
ART	0.127	0.058	0.144	0.052
BIOGRAPHY & AUTOBIOGRAPHY	0.015	0.023	-0.031	0.021
BUSINESS & ECONOMICS	-0.237	0.031	-0.200	0.028
COMICS & GRAPHIC NOVELS	0.037	0.026	0.032	0.023
COOKING	-0.280	0.042	-0.186	0.040
DRAMA	-0.072	0.081	-0.104	0.069
EDUCATION	0.049	0.046	0.026	0.043
FAMILY & RELATIONSHIPS	-0.200	0.074	-0.094	0.070
FICTION: Action & Adventure	0.035	0.157	0.155	0.119
FICTION: Biographical	0.032	0.103	0.105	0.084
FICTION: Coming of Age	-0.159	0.038	-0.035	0.033
FICTION: Crime	0.037	0.040	-0.003	0.036
FICTION: Cultural Heritage	-0.167	0.101	-0.171	0.092
FICTION: Dystopian	-0.100	0.037	-0.006	0.036
FICTION: Family Life	-0.224	0.031	-0.151	0.027
FICTION: Fantasy	-0.086	0.052	-0.030	0.046
FICTION: Friendship	-0.274	0.121	-0.172	0.105
FICTION: General	-0.067	0.036	-0.065	0.031
FICTION: Historical	-0.060	0.048	0.028	0.043
FICTION: Horror	0.262	0.141	0.145	0.117
FICTION: Humorous	-0.264	0.050	-0.162	0.044
FICTION: Literary	-0.177	0.047	-0.019	0.036
FICTION: Mystery & Detective	-0.174	0.034	-0.178	0.029
FICTION: Psychological	-0.174	0.059	-0.180	0.054
FICTION: Religious	-0.087	0.113	-0.166	0.097
FICTION: Romance	-0.126	0.040	-0.050	0.033
FICTION: Science Fiction	-0.107	0.062	-0.006	0.058
FICTION: Thrillers	-0.215	0.034	-0.259	0.030
FICTION: Women	-0.008	0.047	-0.031	0.043
HEALTH & FITNESS	-0.275	0.050	-0.298	0.046
HISTORY	-0.072	0.031	-0.047	0.027
HOUSE & HOME	-0.112	0.044	-0.189	0.041
HUMOR	-0.150	0.037	-0.081	0.036
JUVENILE FICTION	-0.020	0.020	-0.036	0.017
JUVENILE NONFICTION	0.014	0.041	-0.059	0.036
LITERARY COLLECTIONS	-0.112	0.047	-0.082	0.040
LITERARY CRITICISM	-0.095	0.046	-0.063	0.041
MEDICAL	-0.172	0.067	-0.120	0.058
MUSIC	-0.161	0.150	-0.369	0.137
NATURE	-0.241	0.071	-0.127	0.067
PHILOSOPHY	-0.034	0.088	0.137	0.084
POETRY	0.089	0.046	0.062	0.045
POLITICAL SCIENCE	0.061	0.033	0.033	0.031
PSYCHOLOGY	0.071	0.048	-0.119	0.045
RELIGION	-0.015	0.044	-0.047	0.041
SCIENCE	-0.124	0.037	-0.062	0.034
SOCIAL SCIENCE	-0.065	0.034	-0.060	0.029
SPORTS & RECREATION	-0.102	0.060	-0.057	0.056
TRAVEL	-0.293	0.122	-0.488	0.115
UNKNOWN	-0.178	0.042	-0.103	0.035
ATE	-0.082	0.013	-0.069	0.010

Table A.2 Genre-specific sales rank elasticities with respect to star ratings. Standard errors are clustered by edition-platform.

	Panel A: Linear FE		Panel B: DML	
Genre	Estimate	Std. Error	Estimate	Std. Error
ART	-0.197	0.017	-0.195	0.016
BIOGRAPHY & AUTOBIOGRAPHY	-0.176	0.008	-0.165	0.008
BUSINESS & ECONOMICS	-0.169	0.017	-0.160	0.017
COMICS & GRAPHIC NOVELS	-0.144	0.035	-0.543	0.032
COOKING	-0.189	0.050	-0.072	0.045
DRAMA	-0.141	0.092	-0.193	0.067
EDUCATION	-0.196	0.022	-0.217	0.022
FAMILY & RELATIONSHIPS	-0.093	0.022	-0.106	0.022
FICTION: Action & Adventure	-0.141	0.074	-0.162	0.065
FICTION: Biographical	-0.134	0.030	-0.156	0.029
FICTION: Coming of Age	-0.120	0.022	-0.100	0.021
FICTION: Crime	-0.189	0.031	-0.112	0.030
FICTION: Cultural Heritage	-0.085	0.025	-0.046	0.025
FICTION: Dystopian	-0.129	0.036	-0.114	0.036
FICTION: Family Life	-0.192	0.014	-0.208	0.014
FICTION: Fantasy	-0.091	0.021	-0.044	0.020
FICTION: Friendship	-0.182	0.041	-0.170	0.041
FICTION: General	-0.106	0.016	-0.120	0.015
FICTION: Historical	-0.180	0.014	-0.174	0.014
FICTION: Horror	-0.042	0.045	-0.122	0.044
FICTION: Humorous	-0.227	0.023	-0.189	0.023
FICTION: Literary	-0.115	0.041	-0.104	0.042
FICTION: Mystery & Detective	-0.135	0.024	-0.090	0.022
FICTION: Psychological	-0.146	0.022	-0.171	0.020
FICTION: Religious	-0.142	0.045	-0.227	0.042
FICTION: Romance	-0.200	0.023	-0.084	0.021
FICTION: Science Fiction	-0.133	0.035	-0.217	0.034
FICTION: Thrillers	-0.185	0.015	-0.114	0.015
FICTION: Women	-0.196	0.017	-0.149	0.017
HEALTH & FITNESS	-0.253	0.038	-0.138	0.037
HISTORY	-0.236	0.012	-0.209	0.012
HOUSE & HOME	-0.277	0.029	-0.177	0.029
HUMOR	-0.096	0.031	-0.071	0.031
JUVENILE FICTION	-0.131	0.014	-0.113	0.013
JUVENILE NONFICTION	-0.026	0.038	-0.183	0.035
LITERARY COLLECTIONS	-0.191	0.014	-0.199	0.013
LITERARY CRITICISM	-0.100	0.017	-0.178	0.017
MEDICAL	-0.210	0.029	-0.139	0.030
MUSIC	-0.079	0.041	-0.028	0.040
NATURE	-0.210	0.023	-0.176	0.023
PHILOSOPHY	-0.298	0.026	-0.273	0.028
POETRY	-0.090	0.018	-0.122	0.017
POLITICAL SCIENCE	-0.251	0.010	-0.170	0.010
PSYCHOLOGY	-0.155	0.019	-0.131	0.019
RELIGION	-0.219	0.028	-0.270	0.027
SCIENCE	-0.148	0.013	-0.126	0.013
SOCIAL SCIENCE	-0.178	0.010	-0.140	0.011
SPORTS & RECREATION	-0.150	0.025	-0.075	0.026
TRAVEL	-0.103	0.039	-0.065	0.039
UNKNOWN	-0.103	0.029	-0.164	0.025
ATE	-0.173	0.005	-0.156	0.005

Table A.3 Genre-specific sales rank semi-elasticities with respect to newspaper reviews. Standard errors are clustered by edition-platform.

A.2 Genre Merging

Original Genre	Merged Genre
FICTION: Fam...	FICTION: Family Life
FICTION: Cla...	FICTION: General
FICTION: G...	FICTION: General
FICTION: Bed...	FICTION: General
FICTION: Fai...	FICTION: General
FICTION: Native Ameri...	FICTION: Cultural Heritage
FICTION: Act...	FICTION: Action & Adventure
FICTION: 1)	LITERARY CRITICISM
FICTION: 1)LITERARY CRITICISM	LITERARY CRITICISM
FICTION: 5)	LITERARY CRITICISM
FICTION: 5)LITERARY CRITICISM	LITERARY CRITICISM
FICTION: Anthologies ...	LITERARY COLLECTIONS
FICTION: Visionary & Metaphysical	FICTION: Fantasy
FICTION: Occult & Supernatural	FICTION: Fantasy
FICTION: Ghost	FICTION: Fantasy
FICTION: Gothic	FICTION: Horror
FICTION: War & Military	FICTION: Historical
FICTION: Alternative History	FICTION: Science Fiction
FICTION: Christian	FICTION: Religious
FICTION: Jewish	FICTION: Religious
FICTION: LGBT	FICTION: Romance
FICTION: Erotica	FICTION: Romance
FICTION: Animals	FICTION: General
FICTION: Media Tie-In	FICTION: General
FICTION: Political	FICTION: General
ANTIQUES & COLLECTIBLES	ART
ARCHITECTURE	ART
FOREIGN LANGUAGE STUDY	EDUCATION
Computer programming	SCIENCE
COMPUTERS	SCIENCE

Table A.4 Representative examples of 30 out of the 76 total genre-merging rules used to construct the final genre set. The complete mapping is available upon request.

A.3 Confidence Intervals for Differences Between Linear FE and DML Estimates

To assess whether differences in genre-specific CATEs Δ_g are statistically significant, I construct confidence intervals using the following expression:

$$\Delta_g \ \pm \ q_{1-\alpha/2}^{N(0,1)} \frac{\sigma_{\Delta_g}}{\sqrt{N}} \tag{A.1}$$

where $\sigma^2_{\Delta_g} = \mathrm{Var}(\theta_g^{FE}) + \mathrm{Var}(\theta_g^{DML}) - 2\,\mathrm{Cov}(\theta_g^{FE}, \theta_g^{DML})$. Because both θ_g^{FE} and θ_g^{DML} are estimated separately, $\mathrm{Cov}(\theta_g^{FE}, \theta_g^{DML})$ is unknown. The most econometrically sound approach would be to bootstrap the distribution of Δ_g to obtain σ_{Δ_g}. However, this is computationally infeasible because it would require rerunning the full DML procedure on each bootstrap sample. For this reason, I set the covariance term to zero and use the approximation $\sigma^2_{\Delta_g} \approx \mathrm{Var}(\theta_g^{FE}) + \mathrm{Var}(\theta_g^{DML})$. Because the covariance between the two estimators is likely positive, ignoring it results in wider intervals than the true ones. Any significant differences that remain under this conservative approach can therefore be viewed as robust.

A.4 ALE Computation

ALE for feature j at point x is computed by first estimating its uncentered accumulated local effect. Feature x_j is partitioned into a set of disjoint intervals $[z_{k-1}, z_k]$. For each interval k, let $N_j(k)$ denote the set of observations whose values of x_j fall inside that interval, and let $n_j(k) = |N_j(k)|$ be the number of such observations. The uncentered ALE is then:

$$\hat{\hat{f}}_{j,ALE}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{i:x_j^{(i)} \in N_j(k)} \left[f(z_k, x_{-j}^{(i)}) - f(z_{k-1}, x_{-j}^{(i)}) \right] \tag{A.2}$$

where $k_j(x)$ is the index of the interval containing x. The difference $f(z_k, x_{-j}^{(i)}) - f(z_{k-1}, x_{-j}^{(i)})$ represents the local effect of increasing feature x_j from the lower to the upper boundary of interval k for observation i. Averaging these differences within the interval gives the average local effect in that neighborhood. The outer summation over k then accumulates these average effects across all intervals up to the value x. Because this accumulated curve does not have a natural baseline, it is centered so that its mean over all observations is zero:

$$\hat{f}_{j,ALE}(x) = \hat{\hat{f}}_{j,ALE}(x) - \frac{1}{n} \sum_{i=1}^{n} \hat{\hat{f}}_{j,ALE}(x_j^{(i)}) \tag{A.3}$$

where $\hat{\hat{f}}_{j,ALE}(x)$ is the uncentered ALE function, from which its global mean is subtracted. The resulting centered ALE curve can therefore be interpreted as the effect of a specific feature x_j at value x relative to the average prediction level of the ML model in the data (Molnar, 2025).