University of Groningen

Panel Data Analysis

Vomberg, Arnd; Wies, Simone

Published in:
Handbook of Market Research

DOI:
[10.1007/978-3-319-05542-8_19-1](https://doi.org/10.1007/978-3-319-05542-8_19-1)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2021

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Vomberg, A., & Wies, S. (2021). Panel Data Analysis: A Nontechnical Introduction for Marketing Researchers. In C. Homburg, M. Klarmann, & A. E. Vomberg (Eds.), *Handbook of Market Research* Springer. https://doi.org/10.1007/978-3-319-05542-8_19-1

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Panel Data Analysis: A Nontechnical Introduction for Marketing Researchers

Arnd Vomberg and Simone Wies

Contents

Introduction: Relevance of Panel Data for Marketing Research	2
Process for Panel Data Analysis	4
Define the Research Question	5
Collect Panel Data	5
Prepare Panel Data	6
Explore Panel Data	9
Analyze Panel Data Models	16
Interpret and Present Results	34
Additional Methods in Panel Data Analysis	35
Robust Inference	35
Combining the Fixed Effects and Random Effects Estimators	35
Hausman-Taylor Approach: Consistent Estimation of Time-Constant Effects in the Combined Approach	40
Summary of the Discussed Estimators and Their Underlying Assumptions	42
Modeling a Price-Response-Function in Differences	45
Advanced Topics in Panel Data Analysis	47
Dynamic Panel Data Estimation	47
Random Slope Models: A Multilevel Model Approach to Panel Data	51
Addressing Measurement Error with Structural Equation Modeling Based on Panel Data ...	52
Conclusion	54
Cross-References	55
References	55

A. Vomberg (✉)
University of Groningen, Groningen, Niederlande
e-mail: A.E.Vomberg@rug.nl

S. Wies
Goethe University Frankfurt, Frankfurt, Germany
e-mail: wies@econ.uni-frankfurt.de

Abstract

The analysis of panel data is now part of the standard repertoire of marketers and marketing researchers. Compared to the analysis of cross-sectional data, panel data allow marketers to alleviate endogeneity concerns when linking an independent variable (e.g., price) to an outcome variable (e.g., sales volume). The more accurate estimates that result from panel data analysis help improve marketers' decision-making in focal areas such as price setting and marketing budget allocation. Besides, panel data allow marketers to track customer behavior changes and distinguish real loyalty effects (i.e., same customer repeatedly buys a brand) from spurious effects (i.e., the same number of, but each time different set of, customers buys a brand). This chapter provides a nontechnical introduction to panel data analysis. Marketers will learn how to manage and analyze panel datasets in Stata. They will learn about the focal panel data estimators (pooled OLS, fixed effects, and random effects estimator), their underlying assumptions, advantages, and pitfalls. Besides, we introduce the between effects estimator, the combined approach, the Hausman-Taylor approach, and the first differences estimator as further techniques to analyze panel data. Finally, readers will receive an introduction to advanced topics such as dynamic panel models, panel data multilevel modeling, and using panel data to address measurement errors.

Keywords

Cluster-robust standard errors · Dynamic panel data models · Endogeneity · Fixed effects estimator · Hausman test · Hausman-Taylor method · Measurement error · Omitted variable bias · Panel data analysis · Pooled OLS · Random effects estimator · Serial correlation

Introduction: Relevance of Panel Data for Marketing Research

The analysis of panel data is now part of the standard repertoire of marketers and marketing researchers. Panel data, sometimes referred to as longitudinal data, contain observations about different cross-sectional units, also called clusters, across time. Hence, like cross-sectional data, panel data contain observations across a collection of clusters, and like time-series data, panel data contain observations about these clusters repeatedly collected over time. Examples of panel data include the following:

- Retail scanner data: Retailers track sales volume for their products over time.
- Online transaction data: Online retailers collect information about their customers over time.
- Market research institute data: Organizations collect consumer survey data for brands and products over time. For instance, the American Customer Satisfaction Index (ACSI) tracks the evolution of customer satisfaction for different

companies over time (e.g., Fornell et al. 1996). Similarly, the Young&Rubicam Brand Asset Valuator monitors consumers' brand sentiment for brands over time (e.g., Mizik and Pavlov 2021 in this Handbook).

Typically, the collection of panel data requires huge resources in time and money from investigators. Yet, as compared to cross-sectional data, panel data offer substantial advantages that can easily compensate for their data collection efforts.

First, panel data allow to study changes at the individual level and to disentangle real loyalty effects from so-called spurious carryover effects. As an example, aggregated brand sales data might resemble a stable pattern, indicating high loyalty. However, tracking changes at the individual level provides insights into whether the aggregated effect results from a loyalty effect (i.e., the same consumers purchase regularly) or an attraction effect (i.e., the company attracts new consumers but existing consumers do not repurchase).

Second, panel data allow addressing a potential omitted variable bias, a serious endogeneity threat in observational data. Broadly, endogeneity refers to a situation in which an independent variable is correlated with the error term, violating the basic exogeneity assumption of OLS and causing all coefficient estimates of the model to be biased and inconsistent (both properties that lead to misleading hypothesis tests). For instance, the investigator might be interested to know whether a company's distribution intensity contributes to its financial performance. If she omitted company factors such as brand strength, which likely drives both financial performance and access to distribution channels, regression results might overstate the impact of distribution intensity. In panel data analysis, we can control for general company effects, thereby helping to rule out many threats from omitted variables (see also Ebbes et al. 2021 in this Handbook).

Third and relatedly, in contrast to typical applications with cross-sectional data, panel data allow to include time lags between dependent and independent variables. Thereby, panel data open up opportunities for novel research questions and, again, increase the researchers' ability to rule out endogeneity concerns (Bornemann and Hattula 2021, and Valli et al. 2021 in this Handbook further discuss criteria for evaluating causality).

This chapter aims to introduce a general approach to analyzing panel data from an applied perspective, focusing on panel data management and analysis. Figure 1 outlines the process along which we structure our chapter. In the section "[Process for Panel Data Analysis](#)," we discuss the panel data research process step-by-step. Based on a real-life research example, we first define our research objective (Section [Define the Research Question](#)); specifically, we are interested in estimating a price-response-function for a company's newly launched headphone. Then, we discuss the collected dataset, which can be downloaded from the Handbook's Data Appendix (Section [Collect Panel Data](#)), and explain how researchers can prepare the data for the analyses (Section [Prepare Panel Data](#)). Afterward, readers learn how to explore the unique structure in panel data (Section [Explore Panel Data](#)), and we discuss the focal panel data estimators (Section [Analyze Panel Data Models](#)). We conclude this section with a short interpretation of the results (Section [Interpret and](#)

Fig. 1 Panel data research process

[Present Results](#)). In the section “[Additional Methods in Panel Data Analysis](#),” we provide more in-depth background information on panel data estimators. In the section “[Advanced Topics in Panel Data Analysis](#),” we explore more advanced topics in panel data analysis, including dynamic panel models and random coefficient models. Note that we keep a strictly intuitive and applied approach throughout the text but point the reader to the many excellent resources available that provide a more formal treatment of panel data estimation.

Process for Panel Data Analysis

Any research process should start with a clearly defined research question and end with a meaningful discussion of the results. However, a research process that involves panel data considerably differs from a process based on cross-sectional data alone regarding the data management and analysis part. Given that data management and analysis is more intricate for panel data, we devote special attention to describing how to prepare, explore, and analyze panel data.

In the spirit of an applied approach to the panel data research process, we guide the discussion in this chapter along a real-life dataset, using the statistical software

Stata. Although most statistical software packages are well equipped to analyze panel data, Stata is particularly suited for such analysis given its convenient and efficient command structure for panel data.

Define the Research Question

The most important issue relates to the relationships researchers want to investigate. Developing an original and relevant research question is not trivial, yet, as mentioned earlier, panel data can offer exciting opportunities for examining more complex phenomena.

In this chapter, we want to aid a consumer electronics company on its pricing strategy for a newly launched headphone. We are interested in understanding how price-setting influences sales volume. This question is relevant yet challenging for the company as the headphone represents an expansion into a new category, with which the company only has limited experience. Economic theory suggests relying on so-called price-response-functions for price management. Price-response-functions link price to sales volume and represent a prerequisite for price optimization.

Collect Panel Data

Data for estimating price-response-functions can come from various sources, such as customer surveys (Vomberg and Klarmann 2021 in this Handbook), experiments (Eggers et al. 2021, and Klingemann et al. 2021 in this Handbook), or market-level data. In this chapter, we focus on market-level data. The real-life case we follow throughout the chapter is based on a dataset of a medium-sized European consumer electronics company. The company created the dataset to estimate a price-response-function for its most recently launched headphone.

An excerpt of the corresponding dataset available for analysis contains weekly sales data (“sales”) of the headphone model across $n = 19$ of the company’s franchised stores (cross-sectional component, labeled as “storeid”) across a time period of $T = 82$ weeks (time-series component, labeled as “week”), ranging from weeks 37 to 118. The company has recorded the headphone’s retail price (“price”) for each store and week. Besides, the company has collected information about whether the store featured major promotion activities for the headphone in a given week (dummy-coded “promo”: 1 “promotional activity in the focal week” 0 “no promotional activity in the focal week”). Moreover, the company has information on two general store descriptors: whether the store design includes multiple floors (dummy-coded “floor”: 1 “multiple floors” 0 “single floor”) and whether the store is in a premium location with high traffic (dummy-coded “location”: 1 “premium location” 0 “nonpremium location”). The data is initially split across three different data files, mimicking the often disordered situation researchers and companies face when they seek to analyze data.

We begin by reviewing the steps necessary to combine the three files to arrive at the final dataset that we will use in the chapter's remainder. Access to the different datasets is available via the Data Appendix. We provide the three individual raw datasets ("Sales_Original.dta," "Sales_Additional.dta," and "Explanatory.dta"), as well as the final dataset ("Dataset_Final.dta"). "Sales_Original.dta" contains the original headphone sales data the firm collected across its sampled stores. "Sales_Additional.dta" includes observations of an additional store for which data was shared separately. "Explanatory_Long.dta" features the retail price variable as well as the three other explanatory variables (promotional activities, "promo," store floor design, "floor," and store premium location, "location"). The Data Appendix also includes the script (the so-called "do-file" in Stata language) that summarizes the programming commands we use and complements the text with more detailed comments regarding the syntax.

Prepare Panel Data

Our overview begins with looking at three data management challenges that deserve our attention when working with panel datasets. Specifically, we discuss how to transform, combine, and prepare these datasets, so they are ready to use for our statistical analysis.

Transforming the Data Structure: Converting Wide Format to Long Format

In general, there are two ways of organizing panel data: wide format and long format. In wide format, a cluster's repeated measurements are stored in a single row, and each measurement appears in a separate column. Hence, the dataset has as many rows as it has clusters and as many columns as time periods $\times$ number of variables. Many datasets, especially when coming from commercial data providers, store panel data in wide format. Our original sales dataset ("Sales_Original.dta") also comes in wide format, with each row containing sales data for an individual store across all weeks. Using the `list` command, we can see a subset of the first five stores and their respective sales volume in the first ten weeks in the dataset in wide format (Fig. 2).

In contrast, in long format, each measurement occupies one row with as many rows as clusters $\times$ time periods (for balanced data: $n \times T$). The number of columns equals the number of variables. We can convert the data from wide to long format by using the Stata `reshape long` command (Fig. 3).

```
. list storeid sales37-sales46 in 1/5, clean
```

	storeid	sales37	sales38	sales39	sales40	sales41	sales42	sales43	sales44	sales45	sales46
1.	1	309	339	291	208	224	194	181	179	168	312
2.	2	411	277	262	198	205	192	177	167	151	321
3.	3	161	162	161	164	173	166	156	165	144	166
4.	4	235	207	203	195	183	164	197	204	156	189
5.	5	139	139	144	144	141	143	140	63	62	61

Fig. 2 Dataset in wide format

Fig. 3 Converting data from wide to long format

```
. reshape long sales, i(storeid) j(week)
```

Data	wide	->	long
Number of obs.	18	->	1476
Number of variables	83	->	3
j variable (82 values)		->	week
xij variables:			
sales37 sales38 ... sales118		->	sales

Fig. 4 Dataset in long format

```
. list storeid sales in 1/5, clean
```

	storeid	sales
1.	1	309
2.	1	339
3.	1	291
4.	1	208
5.	1	224

The command interprets the sales variables’ suffixes (e.g., sales37, sales38) as denoting the grouping that needs to be expanded to long-form. As visible in the output, we expanded the dataset from 18 observations (18 stores) to 1,476 observations (1,476 store-week pairs). In addition, Stata created a week identifier, called “week,” and collapsed the sales variables (“sales37” – “sales118”) to a new variable, called “sales.” Listing the first five observations of the new long-format dataset (using the `list` command) reveals the following data structure (Fig. 4).

Both datasets have exactly the same information. For panel estimation, however, most statistical software packages, Stata included, require the panel data to be organized in long format, with each observation being a distinct cluster-time pair.

Combining Panel Datasets: Appending and Merging Datasets

Data is often scattered across several data files, and our data example is no exception. It is necessary to combine the individual files into a new file containing all the variables and all the observations needed for analysis. There are different operations through which one can combine panel datasets, and we elaborate on the three most popular operations. First, one can add new *observations* to a given dataset, thereby *appending* data vertically by expanding the number of rows. This operation is typically used when one receives additional data for the same variables and seeks to add these new observations to the first dataset. In our data example, assume the marketing manager responsible for the study was able to add another store to the sample (see “Sales_ Additional.dta”), containing 87 observations. The investigator now needs to append this store’s data to the transformed long-format dataset. The `append` command is simple to use and adds the rows from the second dataset to the end of the first dataset (append using “Sales_ Additional.dta”). This expands the first dataset to 1,563 observations.

Second, one can add new *variables* to a given dataset, thereby *merging* datasets horizontally by expanding the number of columns. This operation is needed when variables are stored across different data files. In our data example, we want to add a set of variables that help explain sales levels. These explanatory variables include the

Fig. 5 Merging datasets

```

.      merge 1:1 storeid week using "Explanatory.dta"

```

Result	# of obs.
not matched	0
matched	1,563 (_merge==3)

retail price, promotional activities, premium location, and floor design and are stored in “Explanatory.dta.” Merging panel datasets requires that both datasets have variables in common on which the merging is based. If matching variables are found, merging two datasets is straightforward using the `merge` command. In our example, “store id” and “week” are the matching variables. We first sort on these matching variables and then perform the merge for a one-to-one (1 : 1) matching (Fig. 5). Note that alternative matching procedures involve one-to-many (1 : m) or many-to-one (m : 1) matching, depending on how the respective datasets are structured. For instance, if we would like to match a dataset that includes zip codes for the stores, we could use the many-to-one matching, knowing that we have multiple stores situated in the same zip code area.

The `merge` command creates a new variable called “`_merge`.” This variable takes a value of 1 if the observation is only contained in the first dataset, a value of 2 if it is included only in the second dataset, and a value of 3 if the observation is present in both datasets. In our example, the observations of the matching variables perfectly match and are present in both datasets. As a result, the “`_merge`” variable takes on the value of 3 for all 1,563 observations. In contrast to the `append` command, the `merge` command does not add new observations to the dataset so the sample size remains unchanged. After we performed the operation, we can delete the “`_merge`” variable (`drop _merge`).

Finally, Stata offers a third type of combining datasets in which one can merge datasets horizontally but form pairwise combinations within-cluster, using the `joinby` command. This command is similar to merging datasets but creates all possible combinations of the observations across both datasets. While not applicable to our data example, it can be a useful command in other settings. For instance, imagine two datasets. The first contains a list of executive team members (i.e., CEO and CMO) across several companies. The second includes a list of awards the companies have received (e.g., Innovative Design Award, Best Place to Work Award). Both datasets contain a company identifier variable, which links the executives and awards that belong to the same company. Using the `joinby` command, we can easily combine the datasets and create a new dataset that includes all combinations of executives and awards, hence one row for each executive and award combination per company.

Preparing the Dataset: Length and Missingness of Panel Data

A final consideration in preparing panel data is to examine the missingness of its observations. When each cross-sectional cluster is observed in each time period, the panel is called *balanced*. Unbalanced panels, in contrast, display missingness in the data so that at least one cluster is not observed in every period. For instance, referring

[illegible]

Fig. 6 Declaring and describing panel datasets

to our example, a store might have joined the study only at a later point in time or might have been closed for some weeks during the observation period for store-remodeling purposes.

The Stata language includes several convenient commands to detect data patterns. Importantly, however, these commands (as do all Stata commands involving panel data) first require the data to be formally declared as panel data by using the `xtset` command. The `xtset` command specifies a cluster identifier (here: store i , labeled “storeid”) and a time identifier (here: week t , labeled “week”). We can access the large array of panel commands belonging to the panel command family (`xt` commands) in Stata with this setup. After having declared our data as panel data, using the `xtset` command, we can inspect missingness patterns with the `xtdescribe` command (Fig. 6).

Studying the missingness patterns over the different stores reveal that the data is not balanced across stores. Roughly 5% of the observations belong to one store, for which we have data after week 118. To simplify the discussion that follows in the chapter, we create a balanced panel dataset that contains only those observations for which we have data for all clusters, that is, only observations reported between weeks 37 and 118. The resulting file will be our final dataset (“Dataset_Final.dta”), with which we will work throughout the remainder of the chapter.

Explore Panel Data

Terminology

Recall that the unique feature of panel data is that these data contain both a cross-sectional component ($i = 1, \dots, n$) and a time component ($t = 1, \dots, T$). The time frame can be regular (e.g., data is collected every week, month, or year) or irregular (e.g., data is collected on specific occasions). Picking up on panel data's dual nature, it is common to classify panel datasets based on the relation between the number of clusters and the number of time periods observed. Panels in which a relatively large number of cross-sectional clusters is tracked over a rather short period of time are

referred to as short or micropanels. For instance, retailers may track customer satisfaction scores for many consumers (cross-sectional component) but only for a few years (time component).

In contrast, panels with frequent measurements for a relatively small section of units vis-à-vis a larger number of time periods are referred to as long or macrounits. For instance, we might track the monthly consumer confidence index back to the 1960s (time component) for only a few EU countries (cross-sectional component). It is worth noting that most datasets in marketing are micropanels and that the more common types of panel estimators are derived under the assumption of relatively fewer time periods and a larger number of cross-sectional units (Wooldridge 2016). Please note that in our example, for illustrative purposes, we work with a reduced dataset that includes only a limited number of stores (cross-sectional units).

Focal Challenge of Panel Data Analysis: Nonindependent Observations

Irrespective of the type of panel dataset, we face one focal challenge in analyzing panel data: repeated measurements of a given cluster over time are nonindependent. This has severe implications for estimating our model, which we discuss next. To illustrate this point, we can think of panel data as a nested data structure. Using an extract from our example, we visualize this idea in Fig. 7. Here, we observe headphone sales for three of the stores at different points in time. The figure illustrates that each sales measurement (lower-level unit) is nested in a particular store measurement (higher-level unit).

The hierarchical nature highlights that individual measurements are likely not independent from each other; rather, sales measurements from one store have more in common (i.e., are more strongly correlated) than they have with sales from different stores. Intuitively, if sales in week 40 were independent of a store’s prior sales, a good prediction would be the mean sales levels for week 39 and week 38 across all stores’ observations. However, it seems that a much better prediction is to rely on the store’s prior measurements. This point can be visualized by plotting the overall mean of sales, also called grand mean, across all measurements and the store-specific means of sales, as shown in Fig. 8. As can be seen in Fig. 8, there is substantial within dependence in the sales data (i.e., measurements within stores are similar).

As a more formal treatment of within dependence, we can use Stata’s time-series lag operators (`L.`) and the `correlate` command (`corr sales L1.sales`) and inspect the serial correlation in sales (see also section “[Pooled OLS Estimator](#)”:

Fig. 7 Illustration of panel structure

Fig. 8 Within and between dependence across stores

Ignoring the Panel Structure” for formal tests of serial correlation). Results suggest that headphone sales correlate substantially ($\rho = 0.67$) over time within a store; hence, there is a high serial correlation present in the sales data, and observations are nonindependent from each other. This dependency, in turn, violates a key assumption of traditional OLS estimation and foreshadows why standard OLS estimation is not feasible for panel data analysis (as detailed in section “Pooled OLS Estimator: Ignoring the Panel Structure”). A thought experiment based on Fig. 7 further underpins this point: if sales measurements within a store are very similar, we do not observe nine observations (three measurements for three stores) but effectively only three. Put differently, it reduces the effective size of the sample we can use in estimating the model, which in turn has implications for calculating the standard errors we use for making inferences. We will revisit the challenge of nonindependent observations in the remainder of the chapter. In doing so, we will shift to the more common terminology of within and between variance instead of within and between dependency. Within variance describes the variation within a cluster across time. Between variance describes the variation between clusters.

Dependent Variable: Between and Within Variance

Given nonindependent observations being the focal challenge in panel data, it is crucial to understand the degree of such dependency. The `xtsum` command provides valuable information on the relative importance of within and between variance in the dependent variable. Figure 9 demonstrates the result for sales ($Sales_{it}$) in our data example.

. xtsum sales						
Variable		Mean	Std. Dev.	Min	Max	Observations
sales	overall	280.1906	131.1669	0	750	N = 1558
	between		110.3184	118.9024	549.5976	n = 19
	within		75.28426	92.88575	695.0443	T = 82

Fig. 9 Decomposing the dependent variable sales volume

Besides providing some general descriptive statistics (including mean and range of the variable, as well as the number of observations, N , number of clusters, n , and number of time periods, t , in the sample), the output distinguishes between an overall, a between, and a within variance in the sales data. The within component focuses on particular clusters (i.e., stores) and describes the data within these clusters (i.e., sales levels from across all periods for an individual store). The between component takes the average per cluster (i.e., average sales level for a particular store) and describes the data based on these averages. The overall component treats the observations as entirely independent and calculates the respective measures without considering their panel nesting.

To support the discussion that follows, we detail how the overall, between, and within variance is calculated. Let

$$\overline{\text{Sales}} = \frac{\sum_{i=1}^n \sum_{t=1}^T \text{Sales}_{it}}{n \times T} \quad (1)$$

be the overall mean of the dependent variable across all observations, also called the grand mean, and

$$\overline{\text{Sales}}_i = \frac{\sum_{t=1}^T \text{Sales}_{it}}{T} \quad (2)$$

be the corresponding within mean for each cluster i (i.e., store); we can easily compute the respective variances.

The overall variance is calculated in the same way as in cross-sectional data without considering any panel structure nesting:

$$\hat{\sigma}_{\text{Sales; Overall}}^2 = \frac{\sum_{i=1}^n \sum_{t=1}^T \left(\text{Sales}_{it} - \overline{\text{Sales}} \right)^2}{n \times T - 1}. \quad (3)$$

To calculate the within variance, we employ the following formula:

$$\hat{\sigma}_{\text{Sales; Within}}^2 = \frac{\sum_{i=1}^n \sum_{t=1}^T (\text{Sales}_{it} - \overline{\text{Sales}}_i)^2}{n \times (T - 1)}. \quad (4)$$

When calculating the within variance, some statistical programs, Stata included, add the grand mean back to the within mean in the numerator $(\text{Sales}_{it} - \overline{\text{Sales}}_i + \overline{\overline{\text{Sales}}})$ to make results comparable across overall, within, and between variance.

The between variance is given as:

$$\hat{\sigma}_{\text{Sales; Between}}^2 = \frac{\sum_{i=1}^n (\overline{\text{Sales}}_i - \overline{\overline{\text{Sales}}})^2}{n - 1} \quad (5)$$

For our sample of 1,558 observations, we estimate a between variance of 110.31 and a within variance of 75.28, indicating that in our sample, the variance between clusters is larger than the variance over time within clusters.

Independent Variables: Time-Constant and Time-Varying Variables

While the dependent variable is, by definition, time-varying in the context of panel data, the independent variables can be time-constant or time-varying. In the following, we will refer to time-constant variables as Z . Those variables constitute cluster characteristics that are stable over the observation period. For instance, in modeling firm sales, McAlister et al. (2016) include the firm's type of business strategy to achieve a competitive advantage as a time-constant independent variable that does not change over the observation period. In our data example, a store's location ($Location_i$) and a store's floor design ($Floor_i$) are both time-constant variables that we include as independent variables in explaining sales.

Other variables, however, may change over time and are therefore called time-varying variables. We will refer to these variables as X in the course of this chapter. Ataman et al. (2010), for example, study brand sales levels and include brand distribution breadth as a time-varying independent variable. Our data example includes retail price ($Price_{it}$) and promotional activities ($Promo_{it}$) as independent variables, which vary over time. We can use Stata's `xtsum` command to confirm the type of variation in the independent variables. Figure 10 shows that $Location_i$ and $Floor_i$ are time-constant as they exhibit zero within variation. $Price_{it}$ and $Promo_{it}$, however, are time-varying and display variation within and between stores.

Since three of the four explanatory variables are binary, a further helpful command is `xttab`, which decomposes categorical variables into within and between variation. Figure 11 shows that, overall, roughly 11% of the store-week sales observations result from stores with a multiple floor store design. The between column repeats the breakdown but does so in terms of stores rather than store-weeks. Given we have a balanced sample with the same number of stores and weeks, and $Floor_i$ varies only between stores, we report the same proportions of

```
. xtsum location floor price promo
```

Variable		Mean	Std. Dev.	Min	Max	Observations	
location	overall	.6842105	.4649788	0	1	N =	1558
	between		.4775669	0	1	n =	19
	within		0	.6842105	.6842105	T =	82
floor	overall	.1052632	.3069907	0	1	N =	1558
	between		.3153018	0	1	n =	19
	within		0	.1052632	.1052632	T =	82
price	overall	55.38318	12.596	29	85	N =	1558
	between		12.40116	38.42683	78.84146	n =	19
	within		3.587608	17.31001	64.3466	T =	82
promo	overall	.2727856	.4455345	0	1	N =	1558
	between		.1266187	.097561	.5487805	n =	19
	within		.4281387	-.2759949	1.175225	T =	82

Fig. 10 Decomposing the independent variables

```
. xttab floor
```

floor	Overall		Between		Within
	Freq.	Percent	Freq.	Percent	Percent
0	1394	89.47	17	89.47	100.00
1	164	10.53	2	10.53	100.00
Total	1558	100.00	19	100.00	100.00

(n = 19)

Fig. 11 Decomposing a time-constant categorical variable

```
. xttab promo
```

promo	Overall		Between		Within
	Freq.	Percent	Freq.	Percent	Percent
0	1133	72.72	19	100.00	72.72
1	425	27.28	19	100.00	27.28
Total	1558	100.00	38	200.00	50.00

(n = 19)

Fig. 12 Decomposing a time-varying categorical variable

observations as in the overall column. Finally, the within percent tells us the fraction of times a store reported a given value of the $Floor_i$ variable. For instance, conditional on a store ever having a $Floor_i$ value of 0 (first line), 100% of its observations have a $Floor_i$ value of 0. These numbers indicate the stability of the variable within a cluster. By definition, a time-constant variable will have a tabulation within-percent of 100, while time-varying variables will have a tabulation within-percent below 100 (Fig. 12).

Fig. 13 Transition probabilities for time-varying categorical variable

xttrans promo			
Promotion	Promotion		Total
	0	1	
0	77.72	22.28	100.00
1	59.95	40.05	100.00
Total	72.90	27.10	100.00

Fig. 14 Transition probabilities for a time-constant categorical variable

xttrans floor			
Multiple floor store	Multiple floor store		Total
	0	1	
0	100.00	0.00	100.00
1	0.00	100.00	100.00
Total	89.47	10.53	100.00

The between-percent informs us about the percentage of stores that have ever reported a given value of the variable, in this case, initiate promotional activities during the observation period. Since all sampled stores had periods in which they engaged in promotional activities ($Promo_{it}$ value of 1) and periods in which they did not engage in promotional activities ($Promo_{it}$ value of 0), the between percentage adds up to 200.

Additional information on within and between variation for categorical variables can also be retrieved through the `xttrans` command, which provides transition probabilities within a cluster from one period to the next. Transition probabilities describe the probabilities of changing the state from one value of a variable to another value.

Figure 13 reports the results for the $Promo_{it}$ variable. As visible from the output, about 40% of the store-week observations that featured promotional activities in one period continue to feature promotional activities in the next period. For those store-week observations that did not feature promotional activities in one period, 22% start doing so in the next period. For time-constant variables, such as $Floor_i$, the diagonal entries will always be 100%, and the off-diagonal entries will always be 0% (Fig. 14).

Finally, it is worth noting that whether the independent variable is time-varying or time-constant has implications for the type of variance we can explain in the dependent variable. Specifically, time-constant variables can only account for between variation in the dependent variable. In contrast, time-varying independent variables can explain both between and within variation in the dependent variable, depending to which extent they vary within and between clusters. For instance, Vomberg et al. (2015) observe that market share tends to be relatively stable over time, while it differs substantially between companies. Hence, this time-varying variable will mainly explain between variation in the dependent variable.

The type of time dependence in the independent variables also helps identify the serial correlation sources in the dependent variable. First, variables that do not change over time (Z) will cause the dependent variable to have similar values across periods. For instance, assume we would like to explain a firm's brand equity levels by its branding strategy (multibrands vs. mono-brand). Since the type of branding strategy a firm pursues is likely to be constant over time (Rao et al. 2004), brand equity levels will be close to each other across periods. In statistical language, this cause of serial correlation is referred to as spurious state dependence. Two measurements (of the dependent variable) may be correlated because they are associated with a further variable (that characterizes the cluster).

Second, time-varying variables (X) might also be a source of serial correlation if they change only slowly over time. For instance, customer satisfaction is likely to drive brand equity. Yet, while customer satisfaction levels can change over time, they might only change to a small amount or just every other year. If customer satisfaction levels influence brand equity levels, the serial correlation among the time-varying customer satisfaction levels leads to statistical dependencies in the brand equity levels. This is another representation of spurious state dependence.

Third, the dependent variable's current value might directly influence the dependent variable's level in the next period. For instance, brand familiarity levels in the current year are likely to impact brand familiarity levels in the next year directly, given multiplier and spillover effects through word-of-mouth, like Lovett et al. (2013) show. This argument provides a substantive and direct cause of serial correlation, referred to as true state dependence.

From this discussion, it follows that if we were able to control for all relevant time-constant (Z) and time-varying (X) independent variables and included a lagged dependent variable (y_{t-1}), measurements of the dependent variable should be independent over time and serial correlation would be zero.

Analyze Panel Data Models

After collecting, preparing, and exploring the panel data structure, we are now ready to turn to the actual analysis of panel data. In this section, we will discuss three estimators that investigators can use to estimate the price-response-function: the (1) pooled OLS (Section [“Pooled OLS Estimator: Ignoring the Panel Structure”](#)), (2) fixed effects (Section [Fixed Effects Estimator](#)), and (3) random effects estimator (Section [Random Effects Estimator](#)). We will discuss the latter two estimators in greater detail since those estimators explicitly account for the panel structure (in contrast to the pooled OLS estimator ignoring the panel structure). Besides, “most panel data models are estimated under either the fixed or random effects assumption” (Verbeek 2017, p. 384). We, therefore, call the fixed effects and random estimators proper panel estimators. To recall, here and in the remainder of the chapter, “panel structure” refers to nonindependent observations grouped by clusters over time. We will also outline the relationships between the estimators (Section [Relationship Between Pooled OLS, Fixed Effects, and Random Effects](#)

Estimators), discuss why they may deliver divergent results (Section [What Do Differences Between Pooled OLS, Fixed Effects, and Random Effects Estimators Imply?](#)), and provide guidance on which estimator should be selected (Section [Hausman Test: Selecting Between the Fixed Effects and the Random Effects Estimator](#)). In section “[Additional Methods in Panel Data Analysis](#),” we will highlight additional methods useful in analyzing panel data. We will point to more advanced panel data analysis strategies in section “[Advanced Topics in Panel Data Analysis](#).”

Pooled OLS Estimator: Ignoring the Panel Structure

We will start our discussion by applying the cross-sectional OLS approach to panel data (Skiera et al. 2021 provide a detailed discussion of OLS regression analysis in this Handbook). This technique is referred to as pooled OLS (POLS). The name reflects that POLS pools all observations across the individual cross-sections and time periods without accounting for the panel structure.

As a first step, we need to decide which variables to include in our regression model. Naturally, we include the impact of the price ($Price_{it}$), as this is our focal variable of interest. Since we rely on observational market-level data (instead of randomized experimental data), we also want to account for relevant control variables that explain variation in the dependent variable and reduce the threat of omitted variables that could bias our results. That is, by including these control variables, we can account for factors that might affect both sales volume and price.

We include an indicator for promotion periods ($Promo_{it}$) which by definition influences price and is also likely to boost sales. We include a variable that distinguishes premium locations ($Location_i$) because location advantages might allow managers to realize higher sales volume and to set higher prices. Finally, we account for multiple floor store design ($Floor_i$). Stores with several floors likely contain a larger sales area and a broader assortment. Consequently, customers face a more appealing shopping experience, leading to higher sales volume (e.g., customers may stay longer in the store and purchase more products) but possibly also higher willingness to pay. Moreover, we account for seasonal differences across weeks by including dummy variables for all but one week; we select the first week in our sample as the reference category. Controlling for the effect of weeks (also called week-fixed effects) picks up market-wide developments that equally impact all stores, for instance, stimulated by online buzz about the company or increased media attention in a focal week.

As a second step, we need to decide on the functional form of the price-response-function. For the sake of simplicity, we will focus on a linear model. However, in general, investigators could also estimate price-response-functions in different forms, such as a multiplicative price-response-function. We formulate the following regression equation, which also contains an error term (ξ , pronounced xi):

$$Sales_{it} = \beta_0 + \beta_1 Price_{it} + \beta_2 Promo_{it} + \beta_3 Location_i + \beta_4 Floor_i + \delta WEEK + \xi_{it} \quad (6)$$

The variable subscript notation confirms that $Price_{it}$ and $Promo_{it}$ are time-varying variables, varying by store i and week t , and that $Location_i$ and $Floor_i$ are time-constant variables that only differ by store i . The bolded expression for weeks indicates that we include a vector of week dummy variables.

POLS relies on the typical cross-sectional OLS assumptions. Thereby, one of the focal assumptions is that the error term does not display serial correlation. As outlined in section “[Independent Variables: Time-Constant and Time-Varying Variables](#),” unless we account for *all* time-varying and time-constant variables that impact the dependent variable, the error term is likely correlated between two measurements of the same cluster.

We will demonstrate that serial correlation persists in our model, even after controlling for price, promotion activities, premium location, multiple floor design, and seasonal week effects. We first predict the regression residuals from Eq. 6 (see Eq. 7):

$$\hat{\xi}_{it} = Sales_{it} - (\beta_0 + \beta_1 Price_{it} + \beta_2 Promo_{it} + \beta_3 Location_i + \beta_4 Floor_i + \delta WEEK) \quad (7)$$

As a measure of serial correlation, we next derive pairwise correlations between different time periods. In Table 1, we compare the serial correlation in the residuals (`predict xi_hat, resid`) of the POLS model (Panel a) and in the raw sales data (Panel b) for 5 selected weeks from our dataset (weeks 37–41) to provide a general intuition. Besides eyeballing across the selected correlations, we can undertake more formal tests of serial correlation, either through the `correlate` command (e.g., `corr xi_hat L1.xi_hat`), as we did in section “[Focal Challenge of Panel Data Analysis: Nonindependent Observations](#),” or through the `regress` command (e.g., `reg xi_hat L1.xi_hat, beta`), as Wooldridge (2010) recommends. Also, Stata offers the user-written `xtserial` command as an alternative way to estimate serial correlation (available from SSC `findit xtserial`). From Table 1, it is easy to see that the serial correlation in the POLS residuals is smaller than the serial correlation in the raw sales data. The inclusion of the independent variables explains the reduction in serial correlation. However, it is also apparent that substantial serial correlation persists even after including the independent variables ($\text{corr}(\xi_t, \xi_{t+1}) = 0.61; p < 0.01$). Variables not included in the model explain the remaining serial correlation.

Serial correlation represents a common problem in POLS estimation, rendering standard errors calculated under the typical OLS assumptions misleading for panel data applications (Verbeek 2017). To partly account for serial correlation, investigators can rely on cluster-robust standard errors. Cluster-robust standard errors are computed differently than common OLS standard errors and account for the fact that the error structure differs across clusters (Section “[Robust Inference](#)” details this point). The Stata suffix `cluster(clustvar)` following the regression command indicates the use of cluster-robust standard errors. Thereby, `clustvar` represents the cross-sectional unit around which we cluster the standard errors. Since we want

Table 1 Pairwise correlations of POLS residuals and raw data between weeks

Panel a)		POLS residual					Raw data					
		ξ_{37}	ξ_{38}	ξ_{39}	ξ_{40}	ξ_{41}	Panel b)	Sales ₃₇	Sales ₃₈	Sales ₃₉	Sales ₄₀	Sales ₄₁
ξ_{37}		1.00					Sales ₃₇	1.00				
ξ_{38}		0.77	1.00				Sales ₃₈	0.79	1.00			
ξ_{39}		0.55	0.79	1.00			Sales ₃₉	0.87	0.86	1.00		
ξ_{40}		0.56	0.89	0.87	1.00		Sales ₄₀	0.71	0.90	0.89	1.00	
ξ_{41}		0.48	0.69	0.96	0.83	1.00	Sales ₄₁	0.74	0.77	0.92	0.92	1.00
Overall corr _{Sample} (ξ_t, ξ_{t+1}) = 0.61						Overall corr _{Sample} (Sales _t , Sales _{t+1}) = 0.67						

Table 2 POLS estimates of the price-response-function

	POLS (OLS standard errors)		POLS (cluster-robust standard errors)	
	<i>Model 1</i>		<i>Model 2</i>	
	B	SE	B	SE
Constant (β_0)	283.75***	(28.64)	283.75***	(89.09)
Price (β_1)	-1.48***	(0.23)	-1.48 ^{n.s.}	(1.57)
Promo (β_2)	47.37***	(7.07)	47.37***	(12.74)
Location (β_3)	132.36***	(6.89)	132.36***	(41.28)
Floor (β_4)	-3.41 ^{n.s.}	(10.50)	-3.41 ^{n.s.}	(31.64)
Week-fixed effects	Included		Included	
R ²	0.35		0.35	

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$; n.s. = not significant

to cluster at the store-level, we use “storeid” as our clustering variable. We estimate the regression model with the following Stata syntax:

```
reg sales price promo location floor i.week, cluster(storeid)
```

In Table 2, we report the POLS estimates. The only difference between the two models is how the standard errors are calculated. Model 1 presents the POLS results with standard errors that follow from the common OLS assumptions. Model 2 relies on cluster-robust standard errors. As in the case of cross-sectional OLS, investigators can evaluate the overall model fit via R^2 values, the overall significance of the model with an F-statistic, and the statistical significance of individual regression coefficients with t-tests.

In our case, the regression results reveal that, in line with economic theory, the price has a negative relationship with sales volume ($\beta_1 = -1.48$). However, the effect of price on sales is only statistically significant in Model 1 ($p < 0.01$). Accounting for serial correlation with cluster-robust standard errors increases the estimated standard errors (Model 2), and in our example, the effect of price on sales turns insignificant. Note that clustering the standard errors does not affect the estimated regression coefficients.

To conclude, applying POLS with cluster-robust standard errors to panel data considers the panel structure to some extent but treats “it a nuisance, not as a phenomenon we are interested in” (Rabe-Hesketh and Skrondal 2012, p. 105). As a result, the challenge of serial correlation remains, and the usefulness of the estimation results is limited. In the following, we will focus on two estimators that more explicitly leverage the panel structure: the fixed effects estimator (Section [Fixed Effects Estimator](#)) and the random effects estimator (Section [Random Effects Estimator](#)).

Modeling the Panel Structure

To introduce the idea of explicitly modeling the panel structure, we depict the development of headphone sales volume over time for five selected stores in Fig. 15. To derive Fig. 15, we fit a set of auxiliary regressions. Specifically, for

Fig. 15 Differences in sales volume development over time (Illustrated for fitted regressions of five selected stores)

each store, we run a regression of sales volume on a time period variable, for now excluding the independent variables we included in the POLS model. We see that some stores have systematically higher or lower sales volume than other stores. These systematic differences result from unobserved factors, captured in the error term ξ_{it} . We can classify these unobserved factors into two categories: time-varying and time-constant unobserved factors. In contrast to the POLS estimator, panel data estimators use this notion and split the error term into two parts: $\xi_{it} = u_i + e_{it}$. Thereby, u_i refers to unobserved predictors of the dependent variables that pertain to the cluster (i.e. store-level). Consequently, they are time-constant. The term e_{it} refers to unobserved predictors of the dependent variable that are time varying. We call ξ_{it} the composite error term, u_i the cluster-specific component, and e_{it} the idiosyncratic error term. Eq. 8 displays the rewritten price-response function:

$$\text{Sales}_{it} = \beta_0 + \beta_1 \text{Price}_{it} + \beta_2 \text{Promo}_{it} + \beta_3 \text{Location}_i + \beta_4 \text{Floor}_i + \delta \text{WEEK} + u_i + e_{it} \quad (8)$$

Figure 16 further illustrates the idea of two error components. In a nutshell, the inclusion of the cluster-specific term u_i extends POLS to handling panel data. The cluster-specific component captures that the mean sales volume per store differs from the overall mean sales volume across all stores. In Fig. 16, Store 1 displays systematically higher sales volume levels while the sales volume of Store 2 is systematically lower than the overall level of sales volume. While the store-specific component is constant over time, the idiosyncratic error term (e_{it}) indicates that

Fig. 16 Decomposition of the composite error term ($\xi_{it} = u_i + e_{it}$)

measurements for a focal store (e.g., $e_{1,37}$) can deviate from this store's mean sales volume (e.g., $\beta_0 + u_1$).

The cluster-specific component captures the effects of all unobserved store-level characteristics (e.g., manager ability). It is referred to as unobserved cluster-level heterogeneity. If the cluster-specific component is positive ($u_i > 0$), the mean composite error term (ξ_{it}) will be positive, leading to larger sales volume levels than predicted by the included independent variables. The reverse holds if the cluster-specific component is negative; in that case, sales levels will be lower than predicted by the included independent variables.

The idiosyncratic error term e_{it} is assumed to have zero population mean and exhibits no correlation both across measurement points and with the independent variables. Depending on the assumptions imposed on u_i , two estimators emerge:

- The fixed-effects estimator, which does not assume uncorrelated unobserved cluster-level heterogeneity: u_i can correlate with the independent variables. Therefore, the fixed effects approach allows consistent estimation even if the investigator omits focal store variables (e.g., manager ability) from the price-response-function. The fixed effects approach estimates the respective cluster-specific effects (u_i).
- The random effects estimator, which assumes uncorrelated cluster-level unobserved heterogeneity: u_i is uncorrelated with the independent variables. Therefore, the random effects approach requires that the investigator includes all focal store variables in the price-response-function. The random-effects approach treats u_i as an unobserved random variable with a particular distribution (e.g., normal distribution). Rather than calculating an estimate for every cluster, it estimates a single variance over the clusters.

Fixed Effects Estimator

The fixed effects estimator accounts for the panel structure by exploiting the within variation in the data. It takes each cluster as its control group and only relies on the cluster's variation to estimate the model. Before we discuss the fixed effects

Table 3 LSDV and fixed effects estimates of the price-response-function

	LSDV		Fixed effects	
	Model 3		Model 4	
	B	SE	B	SE
Constant (β_0)	357.86***	(45.91)	394.66***	(56.13)
Price (β_1)	−1.85*	(1.00)	−1.85*	(1.00)
Promo (β_2)	45.21***	(8.63)	45.21***	(8.57)
Week-fixed effects	Included		Included	
Store-fixed effects	Included		Controlled	
R ²	0.75		0.24	

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$; n.s. = not significant; cluster-robust standard errors (SE) in parentheses

estimator in more detail, we introduce a variant of it, called the least squares dummy variables (LSDV) regression, that shares the same logic but is implemented in a more straightforward way.

Least squares dummy variable regression. The name of the LSDV approach results from its reliance on a set of dummy variables. Specifically, the LSDV approach adds store dummies – referred to as store-fixed effects – for all but one store to our price-response-function (Eq. 9). The left-out store fixed effect serves as the reference category. The LSDV, thus, is the POLS model (Section “[Pooled OLS Estimator: Ignoring the Panel Structure](#)”) with an added vector of store-specific dummy variables.

$$\text{Sales}_{it} = \beta_0 + \beta_1 \text{Price}_{it} + \beta_2 \text{Promo}_{it} + \delta \text{WEEK} + \gamma \text{STORE} + e_{it} \tag{9}$$

We do no longer incorporate u_i in the price-response-function since the store-fixed effects account for all observed and unobserved store-level effects in the data. As a consequence, however, we can no longer include the time-constant variables Location_i and Floor_i . These variables correlate perfectly linearly with the store-fixed effects, and thus their effects cannot be estimated. Time-constant variables that interact with time-varying variables (e.g., $\text{Price}_{it} \times \text{Location}_i$), however, could be included as the interaction does not perfectly correlate with the store-fixed effects.

We can estimate the LSDV model with an OLS regression with store-specific dummy variables (`i.storeid`) and, optionally, cluster-robust standard errors (`cluster(storeid)`). Stata will automatically only include $n-1$ store dummy variables. In section “Robust Inference,” we will elaborate on the rationale of using cluster-robust standard errors in addition to employing a panel estimator.

```
reg sales price promo location floor i.week i.storeid, cluster
(storeid)
```

Model 3 in Table 3 shows the LSDV regression results. Since the LSDV regression relies on the OLS estimator, investigators interpret results in the same manner as POLS estimates (Section “[Pooled OLS Estimator: Ignoring the Panel Structure](#)”). We want to highlight some important aspects:

1. The effect of price is slightly stronger in the LSDV regression than in the POLS model (LSDV: $\beta_1 = -1.85$, $p < 0.10$; POLS: $\beta_1 = -1.48$, *n.s.*) and statistically different from zero. Unobserved store-level variables in the POLS model may have suppressed the effect of price. The inclusion of store-specific effects in the LSDV model picks up these effects and explains this difference.
2. Reported R^2 values of the LSDV are typically large since the store-fixed effects explain all time-constant variation of sales volume between stores. LSDV's R^2 value (Model 3: $R^2 = 0.75$) exceeds the one of the POLS model (Model 1: $R^2 = 0.35$).
3. In line with our discussion, Stata will automatically drop time-constant variables from the model due to perfect collinearity. Even if investigators had included premium location and floor design to the price-response-function, these variables would not have been estimated.
4. The constant (β_0) represents the store-specific intercept of the left-out store dummy variable (i.e., $\text{storeid} = 1$) at the left-out week (i.e., $\text{week} = 37$).
5. Usually, investigators do not report estimates for the store dummy coefficients. They are included for statistical reasons but typically do not convey substantive insights. We can, however, test whether they are jointly significant, using a conventional F-test on the estimated coefficients (`testparm` command in Stata). If the test statistic is sufficiently high, we can reject the null hypothesis of zero store effects.

Within transformation. The LSDV approach estimates one regression parameter per store. In panel data applications with many clusters, the number of dummy variables will become large, reducing the degrees-of-freedom in the model and, therefore, the estimates' precision. An alternative way to estimate the model is to rely on time-demeaned data, which produces the same (or, at least, very close) estimates. The idea is first to calculate cluster-means per store for each variable. In the next step, investigators center each variable on their cluster-means by subtracting the cluster-mean of each variable from its observed values (for a more technical discussion, please refer to Cameron and Trivedi 2005, pp. 726–729; Greene 2003, pp. 194–196; Verbeek 2017, pp. 387–388). In the case of time-constant variables, the cluster-mean equals the observed variable.

$$\begin{aligned}
 (\text{Sales}_{it} - \overline{\text{Sales}}_i) &= (\beta_0 - \beta_0) + \beta_1 (\text{Price}_{it} - \overline{\text{Price}}_i) + \beta_2 (\text{Promo}_{it} - \overline{\text{Promo}}_i) \\
 &\quad + \beta_3 (\text{Location}_i - \text{Location}_i) + \beta_4 (\text{Floor}_i - \text{Floor}_i) + (u_i - u_i) + (e_{it} - \bar{e}_i) \\
 &= \beta_0 + \beta_1 (\text{Price}_{it} - \overline{\text{Price}}_i) + \beta_2 (\text{Promo}_{it} - \overline{\text{Promo}}_i) + (e_{it} - \bar{e}_i)
 \end{aligned} \tag{10}$$

Like in the LSDV approach, time-demeaning removes the model's cluster-specific component (u_i), as visible in Eq. 10. For the same reason, all remaining time-constant variables also disappear (i.e., Location_i and Floor_i). Since time-

demeaning is referred to as within transformation, the fixed effects estimator is also called the within estimator.

To estimate the model, we could manually transform the data and run an OLS regression on the transformed data. However, such manual transformation would lead to calculating the wrong degrees-of-freedom. Importantly, the investigator would need to subtract the number of estimated cluster-means calculated as an intermediate step. Statistical software packages directly correct for this adjustment in the degrees-of-freedom, and hence we recommend relying on these pre-programmed commands. We use Stata's `xtreg` command with the `fe` option, and additionally clustering the standard errors.

```
xtreg sales price promo i.week, fe cluster(storeid)
```

Model 4 in Table 3 displays the results of the fixed effects estimation. The coefficient estimates of $Price_{it}$ and $Promo_{it}$ are identical between the two models. However, the reported R^2 values differ. The reason is that Stata uses a different denominator to compute the R^2 statistic between Model 3 and Model 4. Specifically, while Model 3 is based on the overall variation of sales volume, Model 4 relies only on the within variation of sales volume that enters the R^2 formula (Wooldridge 2016, pp. 437–438).

At this point, it is also worthwhile to review the three different types of R^2 values that are calculated after a fixed effects model. Statistical programs, such as Stata, conventionally report a within- R^2 , a between- R^2 , and an overall- R^2 . All three values provide insights into the model, but the within-value is typically of main interest after a fixed effects estimation. It indicates how much of the variation in the dependent variable within a store is captured by the model. The between- R^2 , correspondingly, describes how much of the variation in the dependent variable between stores is explained by the model. The overall- R^2 is the weighted average of the two.

Finally, the reported constant terms differ between Model 3 and Model 4. Eq. 10 suggests that time-demeaning removes the constant term. However, most statistical software packages do report a constant term by relying on a slightly different within-transformation. Specifically, Stata additionally subtracts each variable's overall mean, the grand mean, from the observed values (Eq. 11). The estimated constant becomes the average of all store-specific effects: $\hat{\beta}_0 = \sum_{i=1}^n \frac{\hat{u}_i}{n} = \bar{u}$. Thus, the interpretation of the constant in Model 4 differs from Model 3.

$$\begin{aligned} \left(Sales_{it} - \left(\overline{Sales}_i - \overline{Sales} \right) \right) &= \beta_0 + \beta_1 \left(Price_{it} - \left(\overline{Price}_i - \overline{Price} \right) \right) \\ &+ \beta_2 \left(Promo_{it} - \left(\overline{Promo}_i - \overline{Promo} \right) \right) + (e_{it} - (\bar{e}_i - \bar{e})) \end{aligned} \quad (11)$$

Random Effects Estimator

The random effects estimator represents another estimator that explicitly considers the panel structure. While the fixed effects estimator controls for the panel structure by removing the model's cluster-specific effects, the random effects estimator directly models the serial correlation stemming from the cluster-specific effects u_i ($\text{corr}(\xi_{it}, \xi_{it-1})$). Under the random effects assumptions, Eq. 12 expresses the serial correlation (Andreß et al. 2013, pp. 77–78 formally derive this equation) as:

$$\text{Corr}(\xi_{it}, \xi_{it-1}) = \frac{\sigma_{u_i}^2}{\sigma_{u_i}^2 + \sigma_{e_{it}}^2} \quad (12)$$

Equation 12 is referred to as the intraclass correlation coefficient (see Misangyi et al. 2006 for an application of the intraclass correlation-coefficient analysis on business segments and firm performance). It relates the cluster-specific variance to the overall variance (sum of variance between clusters and across time). If there were no serial correlation present in the data, Eq. 12 would produce a quantity of zero, and POLS would be an adequate estimator to use. However, while POLS is valid only in the particular case in which serial correlation is zero, the random-effects model is more general and explicitly models the degree of serial correlation.

Researchers can estimate the random effects model with the (feasible) generalized least squares estimator (FGLS). The FGLS estimator is a weighted least square estimator, which attributes more or less weight to a given observation depending on its variance structure (reflecting the relationship between $\sigma_{u_i}^2$ and $\sigma_{e_{it}}^2$). Like the fixed effects estimator, the FGLS estimator applies a quasi-demeaning procedure and subtracts the variable's cluster-mean from the variable's observed values. However, in contrast to the fixed effects estimator, the FGLS estimator only subtracts a fraction (θ , pronounced theta) between 0 and 1 of the cluster-mean from the respective value. Based on the quasi-demeaned data, the OLS estimator can then be applied. Since the FGLS estimator only subtracts a fraction, time-constant variables do not drop out of the model (see Eq. 13). For a more formal discussion of the underlying algebra, we refer to Cameron and Trivedi (2005, pp. 734–736), Greene (2003, pp. 200–205), Verbeek (2017, pp. 391–392), and Wooldridge (2010, Chapter 10).

$$\begin{aligned} (\text{Sales}_{it} - \theta \times \overline{\text{Sales}_i}) &= \beta_0 + \beta_1 (\text{Price}_{it} - \theta \times \overline{\text{Price}_i}) + \beta_2 (\text{Promo}_{it} - \theta \times \overline{\text{Promo}_i}) \\ &\quad + \beta_3 (\text{Location}_i - \theta \times \overline{\text{Location}_i}) + \beta_4 (\text{Floor}_i - \theta \times \overline{\text{Floor}_i}) \\ &\quad + (u_i - \theta \times \overline{u_i}) + (e_{it} - \theta \times \overline{e_i}) \end{aligned} \quad (13)$$

Equation 14 shows how to calculate θ . Essentially, θ comprises the relationship between the residual variance ($\sigma_{e_{it}}^2$) and cluster-specific variance ($\sigma_{u_i}^2$), which researchers can estimate with the FGLS estimator (Verbeek 2017, p. 392; Wooldridge 2016, p. 442).

Table 4 Random effects estimates of the Price-response-function

	Random effects	
	Model 5	
	B	SE
Constant (β_0)	302.46***	(43.78)
Price (β_1)	−1.82**	(0.89)
Promo (β_2)	45.21***	(8.59)
Location (β_3)	132.79***	(41.76)
Floor (β_4)	−1.48 ^{n.s.}	(34.74)
Week-fixed effects	Included	
R ²	0.35	

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$; n.s. = not significant; cluster-robust standard errors (SE) in parentheses; store effects controlled for through random intercepts

$$\theta = 1 - \sqrt{\frac{\sigma^2_{e_{it}}}{\sigma^2_{e_{it}} + T \times \sigma^2_{u_i}}}$$

(14)

We can obtain the random effects estimator using the Stata `xtreg, re` command, and, if we wish to, cluster the standard errors. To report the fraction of θ that Stata subtracts, we include (`theta`) as an additional option.

```
xtreg sales price promo location floor i.week, re cluster
(storeid) theta
```

Table 4 summarizes the results. Investigators can evaluate the overall model fit with the R^2 statistic and the overall model’s statistical significance with a Wald test (instead of an F-test that cannot account for serial correlation in the error term). As with the fixed effects estimator, Stata reports three types of R^2 . We inspect the overall- R^2 for the random effects model. Researchers can test the significance of individual regression coefficients with a z-statistic that draws on a normal distribution. Model 5 demonstrates that the effect of price on sales is negative and statistically significant ($\beta_1 = -1.82$, $p < 0.05$). In addition, promotional activities ($\beta_2 = 45.21$, $p < 0.01$) and stores in more attractive locations realize higher sales volumes ($\beta_3 = 132.79$, $p < 0.01$). However, stores with a multiple floor design ($\beta_4 = -1.48$, *n.s.*) do not associate with higher sales volumes.

The reported θ value (`theta`), which is used for the quasi-demeaning (Eq. 13), is 0.92 for our data. Stata will report only one value for θ if the panel is balanced. If the panel is unbalanced, Stata will report multiple θ values, depending on the number of weeks for which the store is observed.

- We conclude with two additional comments:
1. The random effects approach treats the store-specific effects as unobservable random variables and not as model parameters (as the fixed effects model does). Still, investigators can obtain estimates for the store-specific intercepts

(predict ui, u; Rabe-Hesketh and Skrondal 2012, p. 107, p. 161, discuss further options to get store-specific error terms).

2. As an alternative to the FGLS estimator, investigators can also rely on a maximum likelihood estimator (mle option instead of re) to estimate the random effects model. In general, the two estimators will deliver equivalent results in large samples. We will rely on the maximum likelihood estimator when discussing panel data analysis from a multilevel modeling perspective in the section “Random Slope Models: A Multilevel Model Approach to Panel Data.”

Relationship between Pooled OLS, Fixed Effects, and Random Effects Estimators

Thus far, we relied on three different estimators (POLS, fixed effects, and random effects estimator) to estimate our price-response-function. Table 5 summarizes the results, introducing the two proper panel data estimators first (Models 6 and 7) and keeping the POLS estimator (Model 8) as a benchmark. We note that the three estimators lead to different results. The relationship among the estimators becomes apparent when we study the quasi-demeaning procedure (e.g., $\text{Sales}_{it} - \theta \times \text{Sales}_i$) underlying the random effects estimator in more detail.

Specifically:

- For $\theta = 1$, the random effects estimator becomes the fixed effects estimator.
- For $\theta = 0$, the random effects estimator becomes the POLS estimator.

Table 5 Comparison of fixed effects, random effects, and POLS estimates of the price-response-function

	Fixed effects	Random effects	POLS
	<i>Model 6</i>	<i>Model 7</i>	<i>Model 8</i>
	B (SE)	B (SE)	B (SE)
Constant (β_0)	394.66*** (56.13)	302.46*** (43.78)	283.75*** (89.09)
Price (β_1)	-1.85* (1.00)	-1.82** (0.89)	-1.48 ^{n.s.} (1.57)
Promo (β_2)	45.21*** (8.57)	45.21*** (8.59)	47.37*** (12.74)
Location (β_3)	Omitted	132.79*** (41.76)	132.36*** (41.28)
Floor (β_4)	Omitted	-1.48 (34.74)	-3.41 ^{n.s.} (31.64)
Week-fixed effects	Included	Included	Included
Store-fixed effects	Controlled		
R ²	0.24	0.35	0.35

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$; n.s. = not significant; cluster-robust standard errors (SE) in parentheses; in random effects model store effects controlled for through random intercepts

While in applied research, θ is unlikely to be exactly 0 or 1, the rearrangement of θ in Eq. 15 helps to understand which determinants affect θ .

$$\theta = 1 - \sqrt{\frac{1}{1 + T \times \left(\frac{\sigma_{u_i}^2}{\sigma_{e_{it}}^2} \right)}} \quad (15)$$

First, with an increasing amount of measurement occasions T , the random effects estimator converges to the fixed effects estimator. In our example, we track sales volume over $T = 82$ weeks. The long observation period explains why the random effects estimator produces results close to those of the fixed effects estimator (Stata used $\theta = 0.92$, close to 1). The more measurement points we collect per cluster, the more the time effect dominates the store effect. Notably, the random effect estimator converges to the fixed effects estimator even if omitted time-constant variables correlate with the independent variables.

Second, the more the idiosyncratic error component dominates the store-specific component (i.e., $\frac{\sigma_{u_i}^2}{\sigma_{e_{it}}^2}$ becomes small), the less important the store-specific effect (u_i) becomes. As a consequence, the more negligible difference it makes to explicitly model the panel structure, the more results from the random effects estimator resemble those of the POLS estimator. In the extreme case in which the store-specific variance is zero ($\sigma_u^2 = 0$), the POLS, random effects, and fixed effects estimators produce identical results. However, Wooldridge (2016, p. 443) notes that it is more common that $\frac{\sigma_u^2}{\sigma_e^2}$ is large and that θ will be closer to 1.

Investigators can rely on the Lagrange multiplier test to formally test the null hypothesis of the store-specific effect's variance being zero ($H_0: \sigma_u^2 = 0$), equivalent to testing whether u_i is zero (Rabe-Hesketh and Skrondal 2012). The `xttest0` command in Stata implements the Lagrange multiplier test after the regression command.

```
xtreg sales, re cluster(storeid)
xttest0
```

In our example, the test statistic clearly rejects the null hypothesis ($\chi^2(\text{d.f.} = 1) = 28,030.42$; $p < 0.01$), and we conclude that the store-specific component is different from zero.

What Do Differences Between Pooled OLS, Fixed Effects, and Random Effects Estimators Imply?

A natural follow-up question is as follows: Which estimator should be selected? To answer this question, we have to recall why the three estimators produce divergent estimates. Although all estimators rely on the same dataset, they exploit variation in the data in different ways. The fixed effects estimator purely uses within variation, whereas the POLS and the random effects estimators rely on both sources of variation.

Fig. 17 Comparison of between and within effects

Using both sources of variation, POLS and random effects estimators can thus exploit more variation (i.e., more information) than the fixed effects estimator. As such, POLS and the random effects estimators are more efficient (i.e., producing smaller standard errors) than the fixed effects estimator, with the random effects estimator being the most efficient (Verbeek 2017). Efficiency is a desirable property. The estimated coefficients from each sample draw deviate from the “true” parameters in the population due to sampling error (also referred to as standard error). As we typically only have one sample to use in our analysis, we are interested in the estimator that produces the smallest standard errors, i.e., the most efficient estimator.

However, as a downside, benefitting from those efficiency gains requires us to impose one additional assumption on the POLS and the random effects estimators: All relevant time-constant variables are included in the model. In other words, we assume that the model fully explains the more information that we use in the estimation. This assumption would be met in our example if we could safely assume that $Location_i$ and $Floor_i$ are the only relevant time-constant variables.

However, we might have overlooked other critical time-constant variables. For instance, managers across stores may differ in their levels of experience and capabilities. More experienced and more capable managers might better anticipate future market developments and set more reasonable prices. Since our observation period covers less than 2 years (weeks 37–118), we can assume that those experiences and capabilities are constant over time. Since we did not measure manager ability, it becomes part of the store-specific error term (u_i). This is problematic if manager ability also impacts realized sales volumes, in which case our price-response-function would suffer from endogeneity in the form of an omitted variables bias.

Figure 17 demonstrates two ways in which omitted manager ability could affect our estimated price-response-functions. In this figure, hollow (Store 1) and black-shaded (Store 2) circles represent observed price and sales volume combinations for two stores. Gray-shaded squares are the corresponding store-specific mean values.

Based on these squares, we show the corresponding regression coefficients (β) predicted based on exploiting within (β^{Within}) versus between variation (β^{Between}). If using within versus between variation leads to the same results, then POLS and random effects estimators should be selected as these are the more efficient estimators. If leveraging within versus between variation leads to different results, then the fixed effect estimator should be chosen as estimators based on between variation (i.e., POLS and random effects) are likely to be biased and inconsistent.

Panel a visualizes a situation in which managers in Store 2 have set higher prices than managers in Store 1. However, the resulting sales volume in Store 2 is at the same level as it would be if Store 1 had set the same price (all observed price-sales-volume combinations lie on the same straight line). Consequently, we retrieve the same information when exploiting either the within or between variation. Thus, in Panel a, we obtain the same regression coefficient, irrespective of the type of variation we exploit ($\beta^{\text{Between}} = \beta^{\text{Within}}$). Such a situation is referred to as uncorrelated unobserved cluster-level heterogeneity, unobserved because we do not observe (or measure) manager ability, and uncorrelated because unobserved ability does not impact price (and realized sales volume). In such a situation, POLS and the random effects estimators are consistent and are more efficient than the fixed effects estimator.

Panel b in Fig. 17 demonstrates a situation in which managers in Store 2 have set higher prices than managers in Store 1, yet this time, the observed level of sales volume dropped below the level that Store 1 would have realized. The reason could be that the managers of Store 2 are less capable of inferring demand and hence price setting. These managers might also create less attractive store environments and make less appealing assortment decisions which lower sales volume. Such a situation is referred to as correlated unobserved cluster-level heterogeneity since, in this situation, the omitted capabilities correlate with the price (and the realized sales volume).

The fixed effects estimator is still consistent in such a situation because it only picks up the within variation. However, the between variation suffers from an omitted variable bias. In picking up the between variation, the random effects estimator assumes that the sales differences between the stores are driven by price only. However, sales are correlated with manager ability, which also impacts price setting (i.e., a situation of correlated unobserved cluster-level heterogeneity). In this case, the estimator is based on an invalid assumption. In other words, it uses more information (i.e., between variation in sales) but uses it under a wrong assumption (i.e., pretending that sales volume is only explained by price while sales volume is also explained by other factors that also correlate with price). As a consequence, the POLS and random effects estimators are not consistent anymore. In Panel b), the POLS and random effects estimators would overestimate the impact of price on sales volume ($|\beta^{\text{Between}}| > |\beta^{\text{Within}}|$).

In the following, we introduce a formal test that assists in choosing between the fixed effects and the random effects estimators. Note that we focus on the comparison between the fixed effects and the random effects estimators. If the random effects estimator is consistent, POLS is consistent, yet when choosing between

POLS and random effects, we recommend the more efficient random effects estimator. However, as we will see later (Section [Summary of the Discussed Estimators and Their Underlying Assumptions](#)), the random effects estimator requires additional assumptions compared to POLS. For this reason, we also acknowledge that some econometricians suggest performing POLS despite potential efficiency gains from the random effects estimator (Angrist and Pischke 2008, p. 223).

Hausman Test: Selecting Between the Fixed Effects and the Random Effects Estimator

The trade-off we discussed in the last section raises the natural follow-up question: How do we know when we can rely on the more efficient random effects estimator? Stated differently, how do we know whether the random effects estimator is inconsistent due to omitted time-constant variables (e.g., managerial ability)?

This question requires a formal statistical test. To evaluate whether the random effects estimator is appropriate, we can rely on a Hausman test (Hausman 1978). The basic idea of the Hausman test is that the fixed effects and random effects estimators are consistent under the assumption that the investigator omitted no relevant time-constant variables from the price-response-function. Thus, their estimates should not differ significantly from each other. However, under the assumption that the investigator omitted relevant time-constant variables, only the fixed effects estimator is consistent and the random and fixed effects model estimates should differ significantly (Cameron and Trivedi 2005, p. 717). The Hausman test formally evaluates the null hypothesis of equal fixed effects and random effect estimates ($H_0: \hat{\beta}^{FE} - \hat{\beta}^{RE} = 0$). Investigators can rely on the random effects estimator unless the Hausman test returns a significant test statistic.

For an individual coefficient, the Hausman test can be calculated with the following test statistic (Eq. 16) that follows a χ^2 distribution with 1 degree-of-freedom.

$$\chi^2 = \left(\frac{(\hat{\beta}^{FE} - \hat{\beta}^{RE}) - 0}{SE_{\begin{pmatrix} \hat{\beta}^{FE} & \hat{\beta}^{RE} \end{pmatrix}}} \right)^2 = \left(\frac{(\hat{\beta}^{FE} - \hat{\beta}^{RE}) - 0}{\sqrt{SE_{\hat{\beta}^{FE}}^2 - SE_{\hat{\beta}^{RE}}^2}} \right)^2 \quad (16)$$

If we included $Price_{it}$ as the only independent variable (e.g., xtreg sales price, re) in our price-response-function, we would obtain the random effects ($\hat{\beta}^{RE} = -1.82$; $SE_{\hat{\beta}^{RE}} = 0.53$) and fixed effects ($\hat{\beta}^{FE} = -1.84$; $SE_{\hat{\beta}^{FE}} = 0.52$) estimates that result in a nonsignificant test statistic ($\chi^2(\text{d.f.} = 1) = 0.01$; *n.s.*), favoring the random effects estimator. This result suggests that we have no reason to be concerned about correlated unobserved cluster-specific factors, such as manager ability.

For more common models that include more than one independent variable, we use a generalization of Eq. 16. Based on matrix algebra, we derive the general test statistic of the Hausman test (Eq. 17; Greene 2003, pp. 208–209) as:

$$\chi^2 = \left(\hat{\beta}^{FE} - \hat{\beta}^{RE} \right)' \times \left(\hat{\psi}^{FE} - \hat{\psi}^{RE} \right)^{-1} \times \left(\hat{\beta}^{FE} - \hat{\beta}^{RE} \right) \quad (17)$$

Thereby, $\hat{\beta}^{FE}$ and $\hat{\beta}^{RE}$ include all coefficient estimates of the fixed effects and the random effects model. The matrices $\hat{\psi}^{FE}$ and $\hat{\psi}^{RE}$ include the corresponding estimated variances and covariances of the estimates. The test-statistic follows an asymptotic χ^2 distribution with p degrees-of-freedom, with p being the number of coefficients tested.

The respective Stata syntax easily allows estimating the Hausman test statistic (`hausman`) after running the respective regressions. Note that we choose the `sigmamore` option to prevent Stata from considering the vector of week-fixed effects as tested coefficients and thereby inflate the reported degrees-of-freedom (degrees-of-freedom should be 2, for two tested coefficients, and not 83).

```
xtreg sales price promo i.week, fe
estimates store FE
xtreg sales price promo location floor i.week, re
estimates store RE
hausman FE RE, sigmamore
```

For our data, we obtain an insignificant test statistic ($\chi^2(\text{d.f.} = 2) = 0.09; n.s.$) and conclude that we can trust the random effects results. This result is not surprising since fixed and random effects estimates hardly differ in our example.

Finally, we end our discussion highlighting three important aspects:

1. The standard Hausman test is not valid when the investigator uses cluster-robust standard errors. In section “[Alternative Hausman Test](#),” we will introduce a fully robust Hausman test.
2. Equation 16 serves to demonstrate when the Hausman test is not likely to yield a significant result. First, the test statistic tends to be insignificant when the numerator is small. In this case, the estimates for the $Price_{it}$ coefficients do not differ significantly between the fixed effects and random effects estimators; that is, it does not matter which estimator we use. Second, the Hausman test also becomes insignificant when the denominator is large. If the fixed effects estimator displays a large standard error (e.g., the variation of prices over time is low), the Hausman test likely yields an insignificant result. Third, measurement error can provoke an attenuation bias in the fixed effects estimator, which describes an estimate’s bias converging to zero and underestimating the true value. As a result, fixed effects estimates could be smaller than the estimates from the random effects model, even if no relevant time-constant variables were omitted (Section “[Summary of the Discussed Estimators and Their Underlying](#)”).

Assumptions” and “Addressing Measurement Error with Structural Equation Modeling Based on Panel Data”). Verbeek (2017, p. 395) notes in this regard:

Although the Hausman test is commonly used as a tool to decide between the random effects and the fixed effects estimators, it should be used with caution. Rejection should not automatically be interpreted as evidence that the fixed effects model is appropriate. Conversely, if the Hausman test does not reject it is not necessarily the case that the random-effects model should be preferred.

3. Even though the random effects estimator is biased and inconsistent if important time-constant variables are omitted, this bias is attenuated by the factor $(1-\theta)$ (see the quasi-demeaning in Section [Random Effects Estimator](#)). As a consequence, the bias becomes increasingly less severe the closer θ approaches to 1.

Interpret and Present Results

As the last step, we economically interpret our findings. First, marketing managers are typically interested in the price-elasticity concept: the percentage change in sales volume when there is a 1% increase in price. Formally, the price-elasticity is defined as $\varepsilon = \frac{\partial \text{Sales}}{\partial \text{Price}} \times \frac{\text{Price}}{\text{Sales}}$, where *Price* is price and *Sales* is sales volume. Thereby, the first part is simply the derivative of our price-response-function ($\frac{\partial \text{Sales}}{\partial \text{Price}} = \beta_{\text{Price}}$). For *Price* and *Sales*, we use their respective sample means. Thus, for our example, the price elasticity is -0.36 ($\varepsilon = -1.82 \times \frac{55.38}{280.19}$), a relatively low value when compared to results from a meta-analysis reporting an average price elasticity of -2.62 (Bijmolt et al. 2005). Economically, this finding suggests that marketing managers can expect that a 1% price increase only lowers sales volume by 0.36%. Thus, consumers hardly react to price changes of the newly introduced headphone. Marketing managers, therefore, are likely tempted to increase the price of the headphone.

Based on our price-response-function, managers can also determine the revenue-optimal price. In our case, this price would be 79.95€ ($p^* = \frac{\beta_{\text{Constant}}}{2\beta_{\text{Price}}} = \frac{291}{2 \times 1.82}$), which would imply a price increase of 24.57€ compared to the current average price. Note that for this analysis, we used the intercept ($\beta_{\text{Constant}} = 291$) of a random effects model without week-fixed effects (`xtreg sales price promo location floor, re`); otherwise, the constant term would only be correct for the left-out week (Table 4: $\beta_{\text{Constant}} = 302.46$).

Additional Methods in Panel Data Analysis

Robust Inference

Obtaining correct standard errors of estimators is complicated for panel data since these data are likely to suffer from serial correlation, as mentioned earlier, as well as heteroskedasticity (Cameron and Trivedi 2005). Petersen (2009) compares the performance of different standard errors in panel datasets and, in general, recommends the use of cluster-robust standard errors for micropanel datasets (large n and small T). In the spirit of Petersen's finding, Cameron and Trivedi (2005, p. 725) also recommend to "base inference on [cluster-] robust standard errors that do not require specifying a model for the error correlation." Marketing researchers commonly follow this recommendation (e.g., Bayer et al. 2020, Borah and Tellis 2014; Warren and Sorescu 2017).

Cluster-robust standard errors are calculated based on the observed distribution of residuals from the model. They are robust in the sense that they consider the clustered (or nested) data structure of panel data and assume that observations are independent between clusters but not necessarily within clusters. In other words, cluster-robust standard errors consider each store as a cluster with observations over time, "and arbitrary correlation—serial correlation—and changing variances are allowed within each cluster" (Wooldridge 2016, p. 433). Cluster-robust standard errors are beneficial because they also control for potential heteroskedasticity, a second challenge in estimating standard errors in panel data. Wooldridge (2010, Chapter 10), Greene (2003, pp. 211–213), and Cameron and Trivedi (2005, Chapter 21.2.3) formally derive cluster-robust standard errors in the context of panel data.

We have already discussed the importance of cluster-robust standard errors for POLS (Section "Pooled OLS Estimator: Ignoring the Panel Structure") since serial correlation likely leads to an underestimation of common OLS standard errors. We have also requested cluster-robust standard errors for the fixed effects and random effects models when estimating the price-response-functions. This choice seems reasonable as, in general, accounting for fixed effects or random effects only lowers serial correlation but does not eliminate it. Thus, our general recommendation is to rely on cluster-robust standard errors for panel data analysis.

Combining the Fixed Effects and Random Effects Estimators

Our prior discussion suggested that investigators need to choose between the fixed effects and random effects estimators. However, the literature also offers a combined approach that seeks to leverage the advantages and alleviates the disadvantages of the two estimators. Specifically, it allows including time-constant variables and provides an alternative Hausman test valid for cluster-robust standard errors. First, before we discuss the combined approach, we will introduce an additional estimator – the between effects estimator – which is necessary to understand the combined

Table 6 Comparison of fixed effects, random effects, POLS, and between effects estimates of the price-response-function

	Fixed effects	Random effects	POLS	Between effects
	<i>Model 9</i>	<i>Model 10</i>	<i>Model 11</i>	<i>Model 12</i>
	B (SE)	B (SE)	B (SE)	B (SE)
Constant (β_0)	394.66*** (56.13)	302.46*** (43.78)	283.75*** (89.09)	208.34 ^{n.s.} (245.05)
Price (β_1)	-1.85* (1.00)	-1.82** (0.89)	-1.48 ^{n.s.} (1.57)	-.87 ^{n.s.} (3.18)
Promo (β_2)	45.21*** (8.57)	45.21*** (8.59)	47.37*** (12.74)	119.00 ^{n.s.} (318.37)
Location (β_3)	Omitted	132.79*** (41.76)	132.36*** (41.28)	128.54 ^{n.s.} (57.72)
Floor (β_4)	Omitted	-1.48 ^{n.s.} (34.74)	-3.41 ^{n.s.} (31.64)	-3.76 ^{n.s.} (84.30)
Week-fixed effects	Included	Included	Included	Omitted
Store-fixed effects	Controlled			
R ²	0.24	0.35	0.35	0.26

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$; n.s. = not significant; cluster-robust standard errors (SE) in parentheses; in random effects model store effects controlled for through random intercepts

approach (Section [Between Effects Estimator](#)). Second, we will outline the combined approach (Section [Combined Approach](#)). Third, we will outline the alternative (fully robust) Hausman test that researchers can perform (Section [Alternative Hausman Test](#)). Fourth, we will outline why the combined approach allows the consistent estimation of time-varying variables (Section [Understanding How the Combined Approach Allows Consistent Estimation of Time-Varying Variables](#)).

Between Effects Estimator

Since marketing research rarely applies the between effects estimator (cf. Nath and Mahajan 2008 for an exception, Table 2, Model 3), we will only shortly overview the estimator. The between effects estimator relies exclusively on the between variation in the data and discards all time-series information. As such, the estimator computes

the variables' average values per cluster (e.g., mean price: $\frac{1}{T} \sum_{t=1}^T \text{Price}_{it} = \overline{\text{Price}_i}$;

effective sample size = $n = 19$ stores) and runs an OLS regression on these averages (Eq. 18). Thus, the between effects estimator exploits the information that the fixed effects estimator does not use. In contrast, the fixed effects estimator only exploits variance over time (within store variation; Section [“What Do Differences Between Pooled OLS, Fixed Effects, and Random Effects Estimators Imply?”](#)), for instance, in terms of price and sales volume as in our data example (e.g., $\text{Price}_{it} - \overline{\text{Price}_i}$; effective sample size = $n \times T = 19 \times 82 = 1,558$ observations).

$$\overline{\text{Sales}_i} = \beta_0 + \beta_1 \overline{\text{Price}_i} + \beta_2 \overline{\text{Promo}_i} + \beta_3 \overline{\text{Location}_i} + \beta_4 \overline{\text{Floor}_i} + u_i + \bar{e}_i \quad (18)$$

Equation 18 displays the price-response-function that we will estimate with the between effects estimator. In Stata, investigators can request the between effects estimator with the `be` option after `xtreg`. Since we rely on balanced data, week-fixed effects are the same for all companies and are automatically omitted.

```
xtreg sales price promo location floor, be
```

Model 12 in Table 6 displays the results of the between effects estimator and compares those results to the fixed effects (Model 9), the random effects (Model 10), and the POLS estimators (Model 11). The results of the between effects estimator are not the focus of our discussion. Instead, we use them to extend our discussion from section “What Do Differences Between Pooled OLS, Fixed Effects, and Random Effects Estimators Imply?” and provide further insights into the random effects and POLS estimators.

We see that the random effects (Model 10: $\beta_1 = -1.82, p < 0.05$) and the POLS (Model 11: $\beta_1 = -1.48, n.s.$) estimates of price on sales volume lie in between the fixed effects (Model 9: $\beta_1 = -1.85, p < 0.10$) and between effects estimates (Model 12: $\beta_1 = -0.87, n.s.$). The random effects estimator is closer to the fixed estimator, and the POLS estimator is closer to the between estimator. This general pattern directly follows from our discussion in section “What Do Differences Between Pooled OLS, Fixed Effects, and Random Effects Estimators Imply?”. The POLS and the random effects estimators represent weighted compromises between the fixed effects and the between effects estimators. The random effects estimator is the more efficient one.

In the following, we will demonstrate how we can leverage this idea on the between effects estimator to derive a combined model that allows consistent estimates of time-varying variables (as the fixed effects estimator does) and allows the inclusion of time-constant variables (as the random effects estimator does).

Combined Approach

The combined approach relies on the random effects estimator but adds the cluster-means of the time-varying variables (i.e., store means for $Price_{it}$ and $Promo_{it}$) to the model, just as the between effects estimator does. Eqs. 19 and 20 demonstrate two options of how investigators can employ a combined approach.

$$\begin{aligned} \text{Sales}_{it} = & \beta_0 + \beta_1 \text{Price}_{it} + \beta_2 \text{Promo}_{it} + \beta_3 \text{Location}_i + \beta_4 \text{Floor}_i + \beta_5 \overline{\text{Price}_i} \\ & + \beta_6 \overline{\text{Promo}_i} + u_i + e_{it} \end{aligned} \quad (19)$$

$$\begin{aligned} \text{Sales}_{it} = & \beta_0 + \beta_1 (\text{Price}_{it} - \overline{\text{Price}_i}) + \beta_2 (\text{Promo}_{it} - \overline{\text{Promo}_i}) \\ & + \beta_3 \text{Location}_i + \beta_4 \text{Floor}_i + \beta_5 \overline{\text{Price}_i} + \beta_6 \overline{\text{Promo}_i} + u_i + e_{it} \end{aligned} \quad (20)$$

Including cluster-means of time-varying variables is equivalent to demeaning the raw data. However, given that we apply this procedure only to time-varying variables, we can still retain the time-constant variables Location_i and Floor_i . If we had an unbalanced panel dataset, we would also need to include the time average of any

Table 7 Combined approach estimates of the price-response-function

	Fixed effects	Combined approach (Eq. 19)	Combined approach (Eq. 20)	Between effects
	<i>Model 13</i>	<i>Model 14</i>	<i>Model 15</i>	<i>Model 16</i>
	B (SE)	B (SE)	B (SE)	B (SE)
Constant (β_0)	394.66*** (56.13)	232.82 ^{n.s.} (197.17)	232.82 ^{n.s.} (197.17)	208.34 ^{n.s.} (245.05)
Price (β_1)	-1.85* (1.00)	-1.85* (1.00)	-1.85* (1.00)	-0.87 ^{n.s.} (3.18)
Promo (β_2)	45.21*** (8.57)	45.21*** (8.59)	45.21*** (8.59)	119.00 ^{n.s.} (318.37)
Location (β_3)	Omitted	128.54*** (44.85)	128.54*** (44.85)	128.54 ^{n.s.} (57.72)
Floor (β_4)	Omitted	-3.76 ^{n.s.} (34.50)	-3.76 ^{n.s.} (34.50)	-3.76 ^{n.s.} (84.30)
Mean price (β_5)		.98 ^{n.s.} (2.76)	-0.87 ^{n.s.} (2.65)	
Mean promo (β_6)		73.79 ^{n.s.} (280.36)	119.00 ^{n.s.} (280.56)	
Week-fixed effects	Included	Included	Included	
Store-fixed effects	Controlled			
R ²	0.24	0.35	0.35	0.26

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$; n.s. = not significant; cluster-robust standard errors (SE) in parentheses

time-varying variable. We provide the corresponding Stata syntax in the do-file and report results in Table 7.

In Table 7, Model 14 relies on Eq. 19 and Model 15 on Eq. 20. In both models, the estimated price coefficients (Model 2: $\beta_1 = -1.85$, $p < 0.10$; Model 3: $\beta_1 = -1.85$, $p < 0.10$) are identical to the corresponding fixed effects estimate (Model 1: $\beta_1 = -1.85$, $p < 0.10$). The interpretation of the cluster-mean variables of $Price_{it}$ and $Promo_{it}$, however, differs between Model 14 and Model 15. The cluster-means in Model 15 (e.g., $\beta_5 = -0.87$, n.s.) directly replicate the between effect estimator (Model 16: $\beta_1 = -0.87$, n.s.). Model 14 tests the difference between the fixed effects and the between effects results (e.g., $\beta_5^{\text{Model 2}} = \hat{\beta}_1^{\text{Model 4}} - \hat{\beta}_1^{\text{Model 1}} = -0.87 + 1.85 = 0.98$).

Alternative Hausman Test

The interpretation of Model 14's cluster-means (Table 7) allows an alternative way to test the appropriateness of the random effects estimator. Both the between effects and the fixed effects estimators are consistent when the investigator did not omit any

relevant time-constant independent variables. Significant cluster-means from Model 14 (i.e., significant differences between the fixed and between effect estimates) imply that relevant time-constant variables are omitted. Jointly testing both cluster-means from Model 14 via a Wald test (H_0 : All cluster-mean values are zero; H_0 : $\text{Price}_i = \text{Promo}_i = 0$), thus, represents an alternative to the Hausman test from section “[Hausman Test: Selecting Between the Fixed Effects and the Random Effects Estimator](#).” The test-statistic follows a χ^2 distribution with p degrees-of-freedom, with p being of the number of cluster-means tested. In line with our prior interpretation of the Hausman test, investigators rely on the random effects estimator (i.e., we can drop the cluster-means from Eq. 19) unless the Wald test returns a significant test statistic.

```
test price_mn promo_mn
```

Hausman (1978, p. 1263) and Hausman and Taylor (1981, p. 1382) showed that both tests are asymptotically equivalent (Wooldridge 2010, p. 332). Thus, to test whether omitted time-constant company variables correlate with the independent variables, we can either compare the fixed effects and random effects estimates (which we did in section “[Hausman-Taylor Approach: Consistent Estimation of Time-Constant Effects in the Combined Approach](#)”) or the fixed effects with the between effects estimates which we do in this section. Wooldridge (2010) recommends the latter (regression-based) version using cluster-robust standard errors. For our example, we cannot reject the null hypothesis of equal between effects and fixed effects ($\chi^2(\text{d.f.} = 2) = 0.13$; $p = 0.94$), replicating the substantive conclusion of section “[Hausman Test: Selecting Between the Fixed Effects and the Random Effects Estimator](#).” Thus, we can rely on the random effects estimator.

Understanding How the Combined Approach Allows Consistent Estimation of Time-Varying Variables

The random effects estimator’s core challenge is the assumption that the independent variables are uncorrelated with the unobservable time-constant cluster effects. In the following, we build on Wooldridge (2016, pp. 445–446) to explain why including cluster-means leads to consistent estimates of time-varying independent variables. Let us assume that manager ability was indeed a driver of sales volume, yielding the specification in Eq. 21. For the sake of simplicity, we will assume that we do not need to account for promotion, premium location, and floor design.

$$\text{Sales}_{it} = \beta_0 + \beta_1 \text{Price}_{it} + \beta_2 \text{Manager_Ability}_i + u_i + e_{it} \quad (21)$$

Since we have not collected data on manager ability, we would falsely specify the price-response-function, as in Eq. 22:

$$\text{Sales}_{it} = \beta'_0 + \beta'_1 \text{Price}_{it} + u'_i + e'_{it} \quad (22)$$

The omitted manager ability becomes part of the store-specific error term ($u'_i = \beta_2 \times \text{Manager_Ability} + u_i$). Since manager ability is time-constant in our example, it can only correlate with any time-constant effect in Price_{it} (i.e., $\text{Manager_Ability} = \delta_0 + \delta_1 \times \overline{\text{Price}_i} + \phi_i$). If the cluster-mean of price displays a nonzero relationship with manager ability ($\delta_1 \neq 0$), Price_{it} becomes endogenous, i.e., price correlates with the store-specific error term ($\text{corr}(u_i, \text{Price}_{it}) \neq 0$). However, by including the cluster-mean of Price_{it} , the correlation of Price_{it} and the store-specific error term u_i disappears ($\beta''_0 = (\beta'_0 + \beta_2 \times \delta_0)$, $\beta''_2 = (\beta_2 \times \delta_1)$, and $u'_i = (\beta_2 \times \phi_1 + u_i)$).

$$\begin{aligned} \text{Sales}_{it} &= \beta'_0 + \beta'_1 \text{Price}_{it} + (\beta_2 (\delta_0 + \delta_1 \overline{\text{Price}_i} + \phi_i) + u_i) + e'_{it} \\ &= \beta''_0 + \beta'_1 \text{Price}_{it} + \beta''_2 \overline{\text{Price}_i} + u'_i + e'_{it} \end{aligned} \quad (23)$$

The Price_{it} coefficient estimate (Eq. 23) will be the same as in the fixed effects model. The intuition behind this result is that while the final regression model (Eq. 23) still omits manager ability, it accounts for unobserved correlated ability effects by including the cluster-mean of price. The included cluster-mean of the price typically has no substantive interpretation but serves to control for an omitted variable bias.

Hausman-Taylor Approach: Consistent Estimation of Time-Constant Effects in the Combined Approach

The combined approach (Section [Combining the Fixed Effects and Random Effects Estimators](#)) allows including time-constant variables and allows that the Price_{it} and Promo_{it} variables correlate with the store-specific error component u_i (like the fixed effects estimator). However, the combined model still requires that time-constant variables (e.g., Location_i and Floor_i) and the cluster-specific error component u_i are uncorrelated. If this assumption does not hold, the estimates of time-constant variables are not consistent.

Hausman and Taylor (1981) propose a method for obtaining consistent estimates of time-constant variables. Essentially, the Hausman-Taylor approach treats variables differently, depending on whether they are time-constant, time-varying, and correlated or uncorrelated with the cluster-specific component u_i . Specifically, the approach discriminates between time-varying endogenous, time-varying exogenous variables, as well as time-constant endogenous and time-constant exogenous variables.

For instance, we may have reason to believe that omitted time-constant variables could affect Price_{it} and Floor_i (i.e., these are endogenous variables). Knowing that higher-level company executives and not the individual store managers are involved in promotion timing and location decisions, we are not worried that omitted time-constant store variables could impact Promo_{it} and Location_i (i.e., these are

Table 8 Hausman-Taylor approach estimates of the price-response-function

	Fixed effects	Combined approach (Eq. 19)	Hausman-Taylor
	<i>Model 17</i>	<i>Model 18</i>	<i>Model 19</i>
	B (SE)	B (SE)	B (SE)
Constant (β_0)	394.66*** (56.13)	232.82 ^{n.s.} (197.17)	295.32 ^{n.s.} (306.82)
Price (β_1)	−1.85* (1.00)	−1.85* (1.00)	−1.85* (1.00)
Promo (β_2)	45.21*** (8.57)	45.21*** (8.59)	45.21*** (8.58)
Location (β_3)	Omitted	128.54*** (44.85)	141.43 ^{n.s.} (338.88)
Floor (β_4)	Omitted	−3.76 ^{n.s.} (34.50)	24.49 ^{n.s.} (1008.37)
Mean price (β_5)		0.98 ^{n.s.} (2.76)	
Mean promo (β_6)		73.79 ^{n.s.} (280.36)	
Week-fixed effects	Included	Included	Included
Store-fixed effects	Controlled		
R ²	0.24	0.35	0.34

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$; n.s. = not significant; cluster-robust standard errors (SE) in parentheses

exogenous variables). Please note that these rationales are only exemplary and require more careful theoretical and empirical justification.

The Hausman-Taylor approach’s idea is now to derive so-called panel-internal instrumental variables for the endogenous variables $Price_{it}$ and $Floor_i$. Panel-internal instruments imply that we can use simple transformations of variables that are already included in the price-response-function as instruments. Thus, panel data offer the advantage that investigators do not have to collect external instrumental variables, which are often not readily available.

Specifically, in the Hausman-Taylor approach, demeaned variables (known from the fixed effects estimation) serve as instruments for time-varying endogenous variables (e.g., $(Price_{it} - \overline{Price_i})$ serves as an instrument for $Price_{it}$). Cluster means of time-varying variables serve as instruments for time-constant endogenous variables (e.g., $\overline{Pomo_i}$ serves as an instrument for $Floor_i$). We want to confirm that both instrumental variables are exogenous ($\text{corr}(Price_{it} - \overline{Price_i}, u_i) = 0$ and $\text{corr}(\overline{Pomo_i}, u_i) = 0$). Importantly, the Hausman-Taylor approach requires at least as many time-varying exogenous variables as time-constant endogenous variables for identification. If we had additional time-varying exogenous variables, we could evaluate the strength of the selected instrumental variables using the `xtoverid` command.

We can obtain estimates for the Hausman-Taylor approach with the `xthtaylor` command in which we specify the endogenous variables (both time-varying and time-constant) with the `endog` option; the other variables are considered exogenous. Model 19 in Table 8 displays the results. We see that the Hausman-Taylor approach results in different estimates for some of the variables, including $Floor_i$ and $Location_i$, compared to the combined approach.

```
xthtaylor sales price promo location floor w2-w82, endog(price
floor) vce(cluster storeid)
```

The Hausman-Taylor approach finds initial application in marketing research. For instance, Boulding and Christen (2003, 2008) employ this approach in the context of new product introduction strategies. For further applications and their implied panel-internal instruments, please consult Butt et al. (2018), Germann et al. (2015), Ho-Dac et al. (2013), Rao et al. (2004), and Steenkamp and Geyskens (2014).

Summary of the Discussed Estimators and Their Underlying Assumptions

We have discussed a range of estimators to estimate the price-response-function. The estimators produce different results since they exploit the panel structure in different ways: The between effects estimator uses the between variation in the data while the fixed effects estimator exploits the within variation in the data. Finally, the POLS and the random effects estimators exploit both types of variation, with the latter being more efficient. Notably, if the random effects model's key assumption is met, all potential estimators are consistent, and the random effects estimator is the most efficient estimator. However, if the cluster-specific effects u_i are correlated with the independent variables, the random effects model's key assumption is not met. In this case, only the fixed effects estimator is consistent; the Hausman test (Section "[Hausman Test: Selecting Between the Fixed Effects and the Random Effects Estimator](#)" and "[Alternative Hausman Test](#)") formally tests this assumption.

Given the divergent results, investigators should understand the different assumptions underlying the estimators (called model-identifying assumptions) and justify their choices in applied research. We review those assumptions in Table 9, which we classify into A, B, and C assumptions according to Von Auer (2013). The A-assumptions address the functional specification of the regression model. B-assumptions focus on the error term structure, and C-assumptions relate to individual variables of the model. The following sources provide a more in-depth discussion of these assumptions: Kennedy (2008), Greene (2003), Skiera et al. (2021), Verbeek (2017), Wooldridge (2010, 2016).

Regarding the A-assumptions, all estimators require that the investigator includes all relevant variables in the model (assumption A1). We call such an approach a rich data model. For the POLS and the random effects estimator, this implies including all relevant time-constant and time-varying control variables. The fixed effects

Table 9 Assumptions of the different estimators

	POLS estimator	Fixed effects estimator	Random effects estimator
A1	No relevant time-varying and time-constant variables are missing	No relevant time-varying variables are missing	No relevant time-varying and time-constant variables are missing
A2	True relationship between independent and dependent variable is linear		
A3	Estimated parameters are constant over all observations		
B1	The expected value of the idiosyncratic error term is zero: $E(\xi_{it} \mathbf{X}, \mathbf{Z}) = 0$	The expected value of the idiosyncratic error term is zero: $E(e_{it} \mathbf{X}) = 0$	The expected value of the idiosyncratic error term is zero: $E(e_{it} \mathbf{X}, \mathbf{Z}, u_i) = 0$ The intercept captures the expected value of the unit-specific error term: $E(u_i \mathbf{X}, \mathbf{Z}) = \beta_0$
B2	Homoskedasticity: $\text{Var}(\xi_{it} \mathbf{X}, \mathbf{Z}) = \sigma^2$	Homoskedasticity: $\text{Var}(e_{it} \mathbf{X}) = \sigma_e^2$	Homoskedasticity: $\text{Var}(e_{it} \mathbf{X}, \mathbf{Z}, u_i) = \sigma_e^2$ $\text{Var}(u_i \mathbf{X}, \mathbf{Z}) = \sigma_u^2$
B3	No serial correlation: $\text{Cov}(\xi_{it}, \xi_{is} \mathbf{X}, \mathbf{Z}) = 0$	No serial correlation: $\text{Cov}(e_{it}, e_{is} \mathbf{X}) = 0$	No serial correlation: $\text{Cov}(e_{it}, e_{is} \mathbf{X}, \mathbf{Z}, u_i) = 0$
B4	Error terms are normally distributed: $\xi_{it} \sim N(0, \sigma^2)$	Idiosyncratic error terms are normally distributed: $e_{it} \sim N(0, \sigma_e^2)$	Idiosyncratic and cluster-specific error terms are normally distributed: $e_{it} \sim N(0, \sigma_e^2)$ $u_i \sim N(\text{constant}, \sigma_u^2)$
C1	Error terms are uncorrelated with independent variables: $\text{Cov}(\xi_{it}, \mathbf{X}) = 0$; $\text{Cov}(\xi_{it}, \mathbf{Z}) = 0$	Idiosyncratic error terms are uncorrelated with independent variables: $\text{Cov}(e_{it}, \mathbf{X} - \bar{\mathbf{X}}) = 0$	Idiosyncratic and cluster-specific error terms are uncorrelated with independent variables: $\text{Cov}(e_{it}, \mathbf{X}) = 0$; $\text{Cov}(e_{it}, \mathbf{Z}) = 0$ $\text{Cov}(u_i, \mathbf{X}) = 0$; $\text{Cov}(u_i, \mathbf{Z}) = 0$
C2	No multicollinearity		
C3	Measurement error-free		

Notes: Based on a stylized regression of $y_{it} = \beta_0 + \beta_1 x_{1it} + \dots + \beta_k x_{kit} + \beta_j z_{ji} + \xi_{it}$ with cluster i in time period t . ξ_{it} is the composite error term, u_i is the cluster-specific error component, and e_{it} is the idiosyncratic error term. x_{it} denotes time-varying variables, and z_i indicates time-constant variables. $\mathbf{X}$ represents the vector of time-varying variables x_{it} , and $\mathbf{Z}$ denotes the vector of time-constant variables z_i

estimator only requires that the investigator includes all relevant time-varying control variables in the model.

Assumption A2 relates to the relationship between independent and dependent variables, which we assume is linear. Skiera et al. (2021) discuss different data-transformations appropriate for linearizing curvilinear relationships, and those transformations equally apply to the panel data context.

Assumption A3 involves the effect of independent variables on the dependent variables to be constant. If this is not the case, investigators could include interaction terms to perform a moderated regression analysis (e.g., Vomberg et al. 2015). By including interaction terms, investigators effectively model slope heterogeneity.

While the inclusion of interaction terms requires that researchers measure the respective moderating variables, in section “[Random Slope Models: A Multilevel Model Approach to Panel Data](#),” we will show that panel data methods also model slope heterogeneity without measured moderating variables.

Regarding the B-assumptions, panel data is especially prone to violate assumptions about the error term distribution given the clusters’ dependency. In this regard, we emphasized the need to rely on cluster-robust standard errors if the investigator decides to employ POLS (Section [Pooled OLS Estimator: Ignoring the Panel Structure](#)). We also recommended their usage for the fixed effects and random effects estimators (Section [Robust Inference](#)), as they help account for serial correlation (violation of assumption B3) and heteroskedasticity (violation of assumption B2).

Regarding assumption C1, the POLS estimator assumes that the composite error term does not correlate with the independent variables (exogeneity). The fixed effects estimator requires that the idiosyncratic error term is uncorrelated with the independent variables. Note that because the cluster-specific effects omit all time-constant independent variables from the model, the exogeneity assumption applies to time-varying independent variables only. The exogeneity assumption is also made for the idiosyncratic error term for the random effects estimator, which is assumed to be uncorrelated with the independent variables. Assumption C2 requires that all variables display unique variation and, hence, are not perfectly correlated (no perfect multicollinearity). For the POLS and the random effects estimators, this is fulfilled if variables have unique variance over time, between clusters, or both. For the fixed effects estimator, this assumption requires that variables display unique variation over time. Standard errors of variables with little variation over time will become large and reduce statistical power. This assumption also explains why investigators cannot add time-constant variables in the fixed effects approach because they perfectly correlate with the cluster-specific effect.

In this regard, we emphasize three further comments:

1. Investigators should not decide between the fixed effects and the random effects approach based on whether they are interested in the effect of time-constant variables. If the effect of time-constant variables is of interest to investigators, they may employ the combined approach (Section [Combining the Fixed Effects and Random Effects Estimators](#)) or the Hausman-Taylor approach (Section [Hausman-Taylor Approach: Consistent Estimation of Time-Constant Effects in the Combined Approach](#)).
2. The fixed effects estimator’s inefficiency, which results from little within variation, might favor random effects estimation. For instance, Wolters et al. (2020) justify a random effects over a fixed effects specification by noting a lack of sufficient variation in their focal variable over time. Warren and Sorescu (2017) suggest a random effects approach since their unbalanced panel dataset contains several clusters with only one observation.
3. Moreover, we want to acknowledge recent calls for a more balanced view between bias and efficiency when deciding between the fixed effects and random effects estimator. For instance, Kummer and Schulte (2019) mention little within

variation as a limitation to their fixed effects approach. Additionally, Andreß et al. (2013, p. 173) state the following:

Inefficiency of the [fixed effects] estimator is a particular problem if the within-unit variance is low and variables hardly change over time. Since you are never in the lucky situation of statistical theory, which assumes repeated sampling, your single sample may provide you with estimates quite different from the true population parameters. In that case, the fact that fixed effects are unbiased (i.e., correct on average) is no comfort for you. Hence, more research is needed that provides a more balanced view of both estimators that takes into account both unbiasedness and efficiency.

Finally, all estimators require that the variables are measured without error (assumption C3) as measurement error represents another form of endogeneity that can lead to violation of assumption C1 (Vomberg and Klarmann 2021). In section “Addressing Measurement Error with Structural Equation Modeling Based on Panel Data,” we will demonstrate how investigators can leverage panel data to reduce measurement error concerns. Notably, due to the within-transformation, the fixed effects estimator is particularly susceptible to attenuation bias from measurement error (Angrist and Pischke 2008; Griliches and Hausman 1986; Wooldridge 2016, p. 440).

Modeling a Price-Response-Function in Differences

The discussion so far has focused on modeling the price-response-function in levels. For the sake of completeness, we briefly discuss an additional estimator for the analysis of panel data: the first difference estimator. The first difference estimator takes the first difference of all variables (see Eq. 24) and then performs an OLS regression on the transformed variables (for a more technical discussion, please refer to Cameron and Trivedi 2005, pp. 729–731).

Equation 24 reveals that unobserved time-constant variables disappear after taking the first differences since they do not change over time. Like the fixed effects estimator, the first difference estimator thereby controls for an omitted variable bias stemming from unobserved time-constant variables. As a consequence, the first difference estimator does not allow to include time-constant variables.

Investigators can manually create the first differences of the variables. Alternatively, Stata automatically creates first differences when placing a difference operator (Δ) in front of the respective variable. In line with our prior discussion, we also recommend that researchers should rely on cluster-robust standard errors for first difference models.

Table 10 First difference estimates of the price-response-function

	First difference		Fixed effects	
	<i>Model 20</i>		<i>Model 21</i>	
	B	SE	B	SE
Constant (β_0)	14.79 ^{n.s.}	(20.66)	394.66***	(56.13)
Price (β_1)	-1.66 ^{n.s.}	(1.05)	-1.85*	(1.00)
Promo (β_2)	57.76***	(10.67)	45.21***	(8.57)
Location (β_3)	Omitted		Omitted	
Floor (β_4)	Omitted		Omitted	
Week-fixed effects	Included		Included	
Store-fixed effects	Controlled		Controlled	
R ²	0.23		0.24	

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.10$; n.s. = not significant; cluster-robust standard errors (SE) in parentheses

$$\begin{aligned}
 (\text{Sales}_{it} - \text{Sales}_{it-1}) &= (\beta_0 - \beta_0) + \beta_1(\text{Price}_{it} - \text{Price}_{it-1}) + \beta_2(\text{Promo}_{it} - \text{Promo}_{it-1}) \\
 &\quad + \beta_3(\text{Location}_i - \text{Location}_i) + \beta_4(\text{Floor}_i - \text{Floor}_i) + (u_i - u_i) \\
 &\quad + (e_{it} - \bar{e}_{it-1}) = \beta_1\Delta\text{Price}_{it} + \beta_2\Delta\text{Promo}_{it} + \Delta e_{it}
 \end{aligned} \tag{24}$$

```
reg D.sales D.price D.promo i.week, cluster(storeid)
```

Table 10 demonstrates the results of the analysis. Since the first difference estimator is obtained via OLS, investigators can evaluate model fit with standard measures such as the R^2 statistic and use t-tests to determine the significance of individual regression coefficients. We find a negative, though not significant, impact of price on sales volume (Model 20: $\beta_1 = -1.66$, n.s.). Comparing the first difference estimates with the fixed effects estimates reveals that, despite their similarity in removing time-constant cluster-specific effects, coefficient estimates differ. Only for the two time periods case (e.g., 2 weeks), first-differencing and the within transformation will result in identical results.

Since both estimators are unbiased and consistent, investigators can choose freely between the two estimators. Wooldridge (2016, p. 439) and Cameron and Trivedi (2005, p. 705) though note that under certain conditions (i.e., no serial correlation and a homoskedastic error term structure), the fixed effects estimator is more efficient than the first difference estimator. Relatedly, a drawback of the first difference estimator becomes apparent the more unbalanced the panel data is. In the case of balanced panels, only the first measurement occasion is dropped since investigators cannot calculate the first difference. However, in the case of unbalanced panel data – which represents the typical case in applied research – first differences can tremendously reduce the sample size. Picking up Verbeek's (2017) critique, the investigator might want to carefully inspect the model specification when the fixed effects and first difference estimator yield substantially different results. Such

differences likely point to misspecification issues that might violate the fixed effects estimator's strict exogeneity assumption.

Finally, we want to point to a particular application area of the first difference estimator, common in the marketing literature (e.g., Gill et al. 2017; Manchanda et al. 2015): the difference-in-differences estimator. The difference-in-differences approach mimics an experimental design while using observational data. Its typical set up includes a binary independent variable that discriminates between a treatment and control group (Cameron and Trivedi 2005, Chapter 22; Verbeek 2017, p. 390; Wooldridge 2016, p. 410). Artz and Doering (2021) discuss its application in more detail in this Handbook.

Advanced Topics in Panel Data Analysis

Dynamic Panel Data Estimation

Dynamic Panel Models Without Cluster-Specific Effects

We now consider dynamic panel models in which investigators include the lagged dependent variable as a time-varying variable. These models are also called lagged-response models, autoregressive models, or Markov models, and Eq. 25 shows their more general form.

$$y_{it} = \beta_0 + \lambda_1 y_{i,t-1} + \dots + \lambda_l y_{i,t-l} + \beta_1 x_{1it} + \dots + \beta_k x_{kit} + \xi_{it} \quad (25)$$

The most popular dynamic model is the autoregressive lag-1 (AR1) model, where the current value of the dependent variable (y_{it}) is regressed on its one-period lagged value ($y_{i,t-1}$). Analogously, in an autoregressive lag-2 AR(2) model, the dependent variable is lagged by two periods ($y_{i,t-2}$). Srinivasan (2021) and Wang and Yildirim (2021) offer an in-depth discussion of AR processes in this Handbook, also covering related topics, including vector autoregressive (VAR) models. Note that we assume the independent variables in all AR model variants to be uncorrelated with the error term and the error term to be serially uncorrelated (see Table 9).

Dynamic panel models can be helpful in various situations. First, by including the lagged dependent variable as an independent variable, dynamic panel models help reduce an omitted variable bias. This notion extends our earlier discussion which has focused on accounting for time-constant cluster-specific effects. Including the lagged dependent variable as a time-varying control relaxes the assumption of omitted variables being only time constant and accounts for time-varying effects. In a classical study of explaining market share, Jacobson and Aaker (1985), for instance, use such an approach to model omitted factors such as customer loyalty and distribution systems, which might have influenced market share in the prior periods as well as in the current period. Since no fixed or random effects are included in such a model, it can be conveniently estimated by OLS.

Second, dynamic models are employed when the effect of the lagged dependent variable is itself of scientific interest. For instance, in a study of synergy effects in

multimedia communications, Naik and Raman (2003) examine the degree of carry-over in sales levels.

Dynamic panel models have received quite some attention in the marketing literature. Germann et al. (2015), for instance, offer an extensive study on the presence of a CMO on firm performance using a lagged dependent variable in their model. Homburg et al. (2020) show that a dynamic approach is even feasible when only the dependent variable is constructed as a panel variable. In their study, the authors investigate the impact of multichannel sales system design (obtained from a cross-sectional survey) on firm performance (derived from secondary panel performance data). The authors include a lagged measurement of firm performance as an independent variable to control for variables that equally impacted performance in different time periods.

Despite its advantages, adopting a dynamic panel structure comes with certain limitations that we now briefly review (see Rabe-Hesketh and Skrondal 2012 for a detailed discussion). First, estimating a dynamic panel model is only feasible when occasions are equally spaced in time. For instance, modeling a dynamic panel structure for data collected over several survey waves with different time intervals would not be a sensible task. It is quite a stretch to assume that the lagged dependent variable has the same effect on the current level of the dependent variable regardless of the time interval between them.

Second, the sample size is considerably reduced when a lagged dependent variable is included because lags are missing for each cluster's first observation (Section [Modeling a Price-Response-Function in Differences](#)). In cases of gaps in the data, the problem of missingness becomes exacerbated because the missing measurement itself is discarded as well as the subsequent observation with its missing lagged measurement.

Third, while not necessarily a limitation, it is worth highlighting that the interpretation of the model's coefficients changes when including the lagged dependent variable. The coefficients now describe the independent variable's effect on the difference between the current and lagged dependent variable. Rearranging the AR (1) dynamic panel model, Eq. 26 shows this point clearly. If λ was equal to 1, the equation would model the change in the dependent variable instead of its level.

$$y_{it} - \lambda y_{i,t-1} = \beta_0 + \beta_1 x_{1it} + \dots + \beta_k x_{kit} + \beta_j z_{ji} + \xi_{it} \quad (26)$$

Finally, when including the time-varying lagged dependent variable as a control to account for omitted variables, such a model makes the strong assumption that all within dependence is due to the lagged dependent variable. In other words, the investigator assumes that the omitted variables are fully accounted for by the lagged dependent variable, there are no cluster-specific effects remaining, and there is no serial correlation present in the error term. This is an admittedly strong assumption, which we can test for, and which we relax in the following section in discussing dynamic panel models that also include cluster-specific effects.

Dynamic Panel Models With Cluster-Specific Effects

We now provide a model that accounts for unobserved cluster-level heterogeneity and adds the lagged dependent variable as an independent variable. A useful feature of such a model is that it can distinguish between two competing explanations of within dependence: unobserved cluster-level heterogeneity (represented by the time-constant cluster effects) and state dependence (represented by the lagged dependent variable). For instance, referring to our data example, the within-store dependence of headphone sales might result from some stores employing very capable managers. Capabilities are likely to be constant over the observation period, in which case this type of dependence represents unobserved cluster heterogeneity. Within-store dependence might also result from high current sales attracting high future sales through improved financial resources to attract new sales, a case of true state dependence. The corresponding model takes the following form:

$$\text{Sales}_{it} = \beta_0 + \lambda \text{Sales}_{i,t-1} + \beta_1 \text{Price}_{it} + \beta_2 \text{Promo}_{it} + \beta_3 \text{Location}_i + \beta_4 \text{Floor}_i + \delta \text{WEEK} + u_i + e_{it} \quad (27)$$

It would be tempting to fit the model using the estimation techniques we introduced in the earlier sections. Unfortunately, the conditions for consistent estimation of Eq. 27 are much more demanding than those required for estimations relying on cluster-specific effects or lagged dependent variables alone (Angrist and Pischke 2008).

First, estimating Eq. 27 with the random effects estimator would lead to inconsistent estimates of the coefficients because the lagged dependent variable is per definition correlated with the cluster-specific effect u_i in the error term. As a result, we would run into a problem of correlated unobserved cluster-level heterogeneity and conclude that the random effects estimator is not feasible when estimating dynamic panel models.

Second, estimating Eq. 27 using the fixed effects estimator does not solve this problem either. The within transformation mechanically correlates the within-transformed lagged dependent variable ($\text{Sales}_{i,t-1} - \overline{\text{Sales}_i}$) with the within-transformed error term ($e_{i,t-1} - \bar{e}_i$) since $\text{Sales}_{i,t-1}$ is correlated with $e_{i,t-1}$ and hence with $\bar{e}_i$. Thus, a fixed effects approach is also not feasible when estimating dynamic panel models.

Finally, estimating Eq. 27 using a first difference approach will also produce inconsistent results, as Eq. 28 shows.

$$\begin{aligned} (\text{Sales}_{it} - \text{Sales}_{i,t-1}) &= \lambda (\text{Sales}_{i,t-1} - \text{Sales}_{i,t-2}) + \beta_1 (\text{Price}_{1it} - \text{Price}_{1i,t-1}) \\ &\quad + \beta_2 (\text{Promo}_{it} - \text{Promo}_{i,t-1}) + (u_i - u_i) + (e_{it} - e_{i,t-1}) \end{aligned} \quad (28)$$

The lagged difference of the dependent variable correlates with the difference of the error term ($e_{i,t} - e_{i,t-1}$) because $\text{Sales}_{i,t-1}$ is related to its error term $e_{i,t-1}$. As

such, the first difference approach also violates the assumption of exogeneity (Section [Summary of the Discussed Estimators and Their Underlying Assumptions](#)).

At the same time, note that $\varepsilon_{i,t} - \varepsilon_{i,t-1}$ is not correlated with lagged differences of the dependent variable beyond the first lag ($Sales_{i,t-1}$), opening up the possibility of instrumenting the lagged difference of the dependent variable with higher lags. Under the assumption that the error term is serially uncorrelated, Anderson and Hsiao (1981, 1982) introduced such a panel-internal instrumental-variable approach. They suggest that investigators can either use the second lag of the dependent variable ($y_{i,t-2}$) or the lag of the first difference ($y_{i,t-2} - y_{i,t-3}$) as instrumental variables for the differenced dependent variable ($y_{i,t-1} - y_{i,t-2}$). As with all instrumental variable estimation, such an approach assumes that the instrumental variables fulfill the relevance and validity assumptions (see Ebbes et al. (2021) in this Handbook).

As an application in marketing, Mizik and Jacobson (2008) rely on the Anderson-Hsiao estimator to model changes in firm profitability. Investigators can obtain a version of this estimator via the `xtivreg` command in Stata.

More efficient IV estimators use additional lags of the dependent variable as instruments, an idea Arellano and Bond (1991) and Blundell and Bond (1998) developed in their generalized method of moments estimator. The Arellano-Bond approach identifies how many lags of the dependent variable are valid instruments and includes all of these lags as instruments (together with the first differences of the model's exogenous variables). Wies et al. (2019), for instance, employ the Arellano-Bond approach when modeling the time series of advertising expenses in studying how firms manipulate their advertising efforts in response to receiving shareholder complaints. The Arellano-Bond approach can be implemented in Stata using the `xtabond`, `twostep` command.

The Blundell-Bond approach considers Eqs. 27 and 28 as a system of equations. It then uses the lagged differences of the dependent variable as instruments for the levels equation (Eq. 27) and the lagged levels of the dependent variable as instruments for the differences equation (Eq. 28). The Blundell-Bond estimator has improved precision properties, as reflected in lower standard errors. It is implemented in Stata using the `xtdpdsys` approach or the user-written `xtabond2` command (available from SSC `findit xtabond2`; Roodman 2009), and the syntax structure follows `xtabond`.

Stata's `xtabond` command family provides a set of practical postestimation specification tests to validate the two critical assumptions underlying a panel-internal instrumental variable approach. Specifically, the `estat sargan` command offers a test of overidentifying restrictions, which is useful in confirming the assumption of instrument validity. The `estat abond` command tests whether the error is serially uncorrelated, another desired property we need to confirm when using either the Arellano-Bond or Blundell-Bond estimator. If the test rejects the latter assumption, we can resort to Stata's `xtdpd` command. This command fits a dynamic panel model based on the Arellano-Bond or Blundell-Bond estimator at the cost of a more complicated syntax but, importantly, allows for low-order autocorrelation in the

error term. For a more detailed exposition of handling dynamic panel models in Stata, we recommend Cameron and Trivedi (2005, 2009).

We close with a final recommendation on assessing the usefulness of our selected dynamic panel model when including cluster-specific effects. Instead of estimating variants of the instrumental-variable model (this section), it might be an insightful task to estimate a separate fixed effects (Section [Fixed Effects Estimator](#)) and random effects model (Section [Random Effects Estimator](#)) plus a separate lagged dependent variable model (Section [Dynamic Panel Models Without Cluster-Specific Effects](#)). If the lagged dependent variable model is correct, but one estimates a fixed effects model, the estimated effect will be too large. If a fixed effects model is correct, but one estimates a lagged dependent variable model, the estimated effect will be too small. Therefore, one can think of the fixed effects and the lagged dependent variable models as bounding the true causal effect of interest (Angrist and Pischke 2008, p. 246).

Random Slope Models: A Multilevel Model Approach to Panel Data

In section “[Dependent Variable: Between and Within Variance](#),” we introduced the focal challenge of nonindependent sales observations in panel data and illustrated this point along with a nested structure figure (Fig. 7). As visible in this figure, panel data can be considered a multilevel model (please refer to Haumann et al. 2021 in this Handbook for a general introduction to multilevel modeling).

In multilevel terminology, the panel dataset represents a two-level data structure, i.e., sales volume measured over time (Level 1) nested in stores (Level 2). In contrast to typical multilevel data, such as sales reps (Level 1) nested in sales managers (Level 2), panel data have an inherent order at the lowest level.

Knowledge of multilevel models allows expanding the discussion on the random effects estimator. Specifically, multilevel models can consider random cluster-specific intercepts as well as random cluster-specific slope coefficients. While prior models (Section “[Analyze Panel Data Models](#)” and “[Additional Methods in Panel Data Analysis](#)”) assumed that all regression coefficients (besides the cluster-specific intercepts u_i) are the same across stores, in the following, we will allow divergent slope coefficients across clusters (Hox 2010; Raudenbush and Bryk 2002).

For instance, we can extend the random effects model and allow $Price_{it}$ to vary between stores. We indicate this additional variability by including the subscript i to the focal regression coefficient (Eq. 29). Eqs. 30 and 31 formally describe the random intercept and random slope, respectively. Eq. 32 shows that we chose not to include a random slope for $Promo_{it}$.

$$Sales_{it} = \beta_{0i} + \beta_{1i}Price_{it} + \beta_{2i}Promo_{it} + \delta WEEK + e_{it} \quad (29)$$

$$\beta_{0i} = \gamma_{00} + \gamma_{01}Location_i + \gamma_{02}Floor_i + u_{0i} \quad (30)$$

$$\beta_{1i} = \gamma_{10} + u_{1i} \quad (31)$$

$$\beta_{2i} = \gamma_{20} \quad (32)$$

Substituting Eqs. 30, 31, and 32 into Eq. 29 leads to the complete multilevel regression model (Eq. 33). Equation 33 resembles the random effects model (Section [Random Effects Estimator](#)) and includes the random slope coefficient of the price ($u_{1i} \times \text{Price}_{it}$).

$$\begin{aligned} \text{Sales}_{it} = & (\gamma_{00} + \gamma_1 \text{Location}_i + \gamma_2 \text{Floor}_i + u_{0i}) + (\gamma_{10} + u_{1i}) \text{Price}_{it} + \gamma_{20} \text{Promo}_{it} \\ & + \delta \text{WEEK} + e_{it} = \gamma_{00} + \gamma_{10} \text{Price}_{it} + \gamma_{20} \text{Promo}_{it} + \gamma_{01} \text{Location}_i \\ & + \gamma_{02} \text{Floor}_i + \delta \text{WEEK} + u_{0i} + u_{1i} \times \text{Price}_{it} + e_{it} \end{aligned} \quad (33)$$

Including random slopes represents an extension of the random effects model and requires a different estimator. While FGLS (`xtreg`, `re`) is typically used to estimate random effects models, multilevel models require a maximum likelihood estimator.

Stata users need to employ the `xtmixed` command, whose syntax slightly differs from the `xtreg` command, to estimate random slope models. Rabe-Hesketh and Skrondal (2012) offer an in-depth discussion on how to build multilevel models in Stata. As before, we recommend that investigators rely on robust standard errors (`vce(robust)` option).

```
xtmixed sales price promo location floor i.week || storeid:
price, vce(robust)
```

Marketing researchers commonly perform random slope applications of multilevel modeling on panel data. For instance, Anderson et al. (2004) rely on panel data in their multilevel model on customer satisfaction's impact on firm performance. Overall, the authors observe a positive effect of customer satisfaction. However, they also find that the effect of customer satisfaction on firm performance significantly varies between companies and industries. Gruca and Rego (2005), Sorescu and Spanjol (2008), Vomberg et al. (2015), and Wlömert and Papies (2019) similarly employ multilevel modeling to panel data to obtain deeper insights on how effects vary across clusters.

Addressing Measurement Error with Structural Equation Modeling Based on Panel Data

Our discussion so far has focused on the benefits of panel data to address endogeneity concerns that may arise from an omitted variable bias. In this section, we discuss the implications of another essential source of endogeneity: measurement

Fig. 18 Measurement error in dynamic panel models illustrated

error, that is, “situations where one or more regressors cannot be measured exactly and are observed with an error” (Ebbes et al. 2021, p. 6).

All estimators discussed in the previous sections rely on the assumption that the variables are measured without error (Section [Summary of the Discussed Estimators and Their Underlying Assumptions](#)). However, this assumption might be violated already in the context of rather objectively verifiable information such as reported price or sales volume. Measurement error may arise due to transmission errors into databases. The measurement error problem becomes even more concerning for more abstract constructs frequently investigated in marketing research (e.g., consumers’ brand perceptions or customer satisfaction). In this regard, measurement theory suggests that observed variables (e.g., observed customer satisfaction scores) represent the net result of a true score and some random error (Vomberg and Klarmann 2021).

Figure 18 illustrates the measurement error problem (Andreß et al. 2013). Please assume that we measured customer satisfaction scores for two consecutive years (t and $t + 1$). Following standard conventions in the literature (e.g., Baumgartner and Weijters 2021), we present the observed values of customer satisfaction in rectangular boxes. As predicted by measurement theory, those values are influenced by the true level of customer satisfaction (“Customer Satisfaction*”), which is specified as circled in Fig. 18 and by a measurement error (ϵ_t).

The amount of measurement error will impact the estimated relationship between customer satisfaction over time, which we could model with dynamic panel data models (Section [Dynamic Panel Data Estimation](#)). For instance, true state dependence (serial correlation of customer satisfaction over time) might be $r = 0.82$ ($\text{corr}(\text{CustomerSatisfaction}^*_t; \text{CustomerSatisfaction}^*_{t+1})$). However, we measured customer satisfaction with error so that the true customer satisfaction level is not translated one-to-one into an observable customer satisfaction level. Statistically speaking, customer satisfaction has a factor loading of $\lambda = 0.70$ (if customer satisfaction was measured without error, the factor loading would be $\lambda = 1.00$). As

a consequence, the observed correlation between customer satisfaction becomes smaller ($\text{corr}(CS_t, CS_{t+1}) = 0.70 \times 0.82 \times 0.70 = 0.40$) and true state dependence is underestimated.

In the context of cross-sectional data, problems of measurement error only concern the independent variables. The error term captures the measurement error of the dependent variable. However, in the context of panel data, measurement error concerns the dependent variable, too. For instance, as illustrated previously, measurement error will bias state dependence estimates toward zero in dynamic panel data models (Section [Dynamic Panel Data Estimation](#)). Additionally, measurement error does not only impact models in which lagged dependent variables are of substantive interest. Demeaning (fixed effects estimator), quasi-demeaning (random effects estimator), or first-difference transformations are equally affected by measurement error. The fixed effects estimator is particularly susceptible to an attenuation bias from measurement error (Angrist and Pischke 2008; Griliches and Hausman 1986; Wooldridge 2016, p. 440).

Structural equation modeling represents a way to directly model measurement error (Baumgartner and Weijters 2021 present an introduction to structural equation modeling in this Handbook). In cross-sectional analyses, researchers use different indicators for the same construct to estimate a construct's reliability and to directly model measurement error. In the panel data context, researchers can use the same indicators from different time periods to capture the underlying construct. For instance, the investigator may use customer satisfaction scores of 3 years (e.g., customer satisfaction measured in $t = 2019, 2020$, and 2021) to estimate a latent customer satisfaction construct. This approach is identical to an evaluation of test-retest reliability and offers the advantage of directly accounting for measurement error. However, the downside of this approach is that fewer possibilities exist to estimate the model (e.g., employing fixed effects estimation is not feasible). Investigators can implement a structural equation model via Stata's `sem` and `gsem` commands.

Luo and Bhattacharya (2006) offer an application example in the marketing context. The authors rely on panel data obtained from Fortune's Most Admired Companies and use repeated corporate social responsibility measurements to capture the underlying latent corporate social responsibility construct. Cho and Pucik (2005) apply a similar approach when modeling how firm innovativeness and product quality relate to market value.

Conclusion

This chapter sought to provide a gentle nontechnical introduction to panel data analysis for marketing researchers. At the core of panel data analysis is the challenge of how best to account for the dependency of observations within and across clusters. We discussed the POLS estimator's limitations and reviewed the two most popular panel estimators that explicitly model the panel structure: the fixed effects and random effects estimators. For completeness, we also discussed the between effects

estimator and first difference estimator, as well as the combined approach and the Hausman-Taylor approach. Using a real-life example, we applied these estimators in the context of a price-response-function for headphone sales. We conducted the empirical analysis in Stata, a very user-friendly statistical software package for analyzing panel data. Despite using the same dataset to estimate the price-response-functions, we find results differ considerably depending on the selected estimator. These divergent results demonstrate the need to thoroughly understand the different model-identifying assumptions, benefits, and limitations of each estimator. As such, we hope our chapter contributes to turning readers into cognizant “regression engineers” (Germann et al. 2015) and offers researchers the necessary skill set to conduct meaningful analyses. Panel data provide exciting opportunities to investigate new research questions, and we hope that readers find this introduction helpful in developing their models.

Cross-References

- ▶ [Applied Time Series Analysis](#)
- ▶ [Assessing the Financial Impact of Brand Equity with Short Time-Series Data](#)
- ▶ [Choice-Based Conjoint Analysis](#)
- ▶ [Dealing with Endogeneity: A Nontechnical Guide for Marketing Researchers](#)
- ▶ [Exploiting Data from Field Experiments](#)
- ▶ [Modeling Marketing Dynamics Using Vector Autoregressive \(VAR\) Models](#)
- ▶ [Multilevel Modeling](#)
- ▶ [Regression Analysis](#)
- ▶ [Structural Equation Modeling](#)
- ▶ [Willingness to Pay](#)

References

- Anderson, T. W., & Hsiao, C. (1981). Estimation of dynamic models with error components. *Journal of the American Statistical Association*, 76(375), 598–606.
- Anderson, T. W., & Hsiao, C. (1982). Formulation and estimation of dynamic models using panel data. *Journal of Econometrics*, 18(1), 47–82.
- Anderson, E. W., Fornell, C., & Mazvancheryl, S. K. (2004). Customer satisfaction and shareholder value. *Journal of Marketing*, 68(4), 172–185.
- Andreß, H.-J., Golsch, K., & Schmidt, A. W. (2013). *Applied panel data analysis for economic and social surveys*. Berlin, Heidelberg: Springer-Verlag.
- Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. New Jersey: Princeton University Press.
- Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies*, 58(2), 277–297.
- Artz, M., & Doering, H. (2021). Exploiting data from field experiments. In C. Homburg, M. Klarmann, & A. Vomberg (Eds.), *Handbook of market research*. Cham: Springer.
- Ataman, M. B., Van Heerde, H. J., & Mela, C. F. (2010). The long-term effect of marketing strategy on brand sales. *Journal of Marketing Research*, 47(5), 866–882.

- Baumgartner, H., & Weijters, B. (2021). Structural equation modeling. In C. Homburg, M. Klarmann, & A. Vomberg (Eds.), *Handbook of market research*. Cham: Springer.
- Bayer, E., Srinivasan, S., Riedl, E. J., & Skiera, B. (2020). The impact of online display advertising and paid search advertising relative to offline advertising on firm performance and firm value. *International Journal of Research in Marketing*, 37(4), 789–804.
- Bijmolt, T. H., Van Heerde, H. J., & Pieters, R. G. (2005). New empirical generalizations on the determinants of price elasticity. *Journal of Marketing Research*, 42(2), 141–156.
- Blundell, R., & Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics*, 87(1), 115–143.
- Bornemann, T., & Hattula, S. (2021). Experiments in market research. In C. Homburg, M. Klarmann, & A. Vomberg (Eds.), *Handbook of market research*. Cham: Springer.
- Boulding, W., & Christen, M. (2003). Sustainable pioneering advantage? Profit implications of market entry order. *Marketing Science*, 22(3), 371–392.
- Boulding, W., & Christen, M. (2008). Disentangling pioneering cost advantages and disadvantages. *Marketing Science*, 27(4), 699–716.
- Borah, A., & Tellis, G. J. (2014). Make, buy, or ally? Choice of and payoff from announcements of alternate strategies for innovations. *Marketing Science*, 33(1), 114–133.
- Butt, M. N., Antia, K. D., Murtha, B. R., & Kashyap, V. (2018). Clustering, knowledge sharing, and intrabrand competition: A multiyear analysis of an evolving franchise system. *Journal of Marketing*, 82(1), 74–92.
- Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. New York: Cambridge University Press.
- Cameron, A. C., & Trivedi, P. K. (2009). *Microeconometrics using Stata*. Texas: Stata Press.
- Cho, H. J., & Pucik, V. (2005). Relationship between innovativeness, quality, growth, profitability, and market value. *Strategic Management Journal*, 26(6), 555–575.
- Ebbes, P., Papies, D., & van Heerde, H. J. (2021). Dealing with endogeneity: A nontechnical guide for marketing researchers. In C. Homburg, M. Klarmann, & A. Vomberg (Eds.), *Handbook of market research*. Cham: Springer.
- Eggers, F., Sattler, H., Teichert, T., & Völckner, F. (2021). Choice-based conjoint analysis. In C. Homburg, M. Klarmann, & A. Vomberg (Eds.), *Handbook of market research*. Cham: Springer.
- Fornell, C., Johnson, M. D., Anderson, E. W., Cha, J., & Bryant, B. E. (1996). The American customer satisfaction index: Nature, purpose, and findings. *Journal of Marketing*, 60(4), 7–18.
- Germann, F., Ebbes, P., & Grewal, R. (2015). The chief marketing officer matters! *Journal of Marketing*, 79(3), 1–22.
- Gill, M., Sridhar, S., & Grewal, R. (2017). Return on engagement initiatives: A study of a business-to-business mobile app. *Journal of Marketing*, 81(4), 45–66.
- Greene, W. H. (2003). *Econometric analysis* (6th ed.). United Kingdom: Pearson Education.
- Griliches, Z., & Hausman, J. A. (1986). Errors in variables in panel data. *Journal of Econometrics*, 31(1), 93–118.
- Gruca, T. S., & Rego, L. L. (2005). Customer satisfaction, cash flow, and shareholder value. *Journal of Marketing*, 69(3), 115–130.
- Haumann, T., Kassemeier, R., & Wieseke, J. (2021). Multilevel modeling. In C. Homburg, M. Klarmann, & A. Vomberg (Eds.), *Handbook of market research*. Cham: Springer.
- Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the Econometric Society*, 1251–1271.
- Hausman, J. A., & Taylor, W. E. (1981). Panel data and unobservable individual effects. *Econometrica: Journal of the Econometric Society*, 1377–1398.
- Ho-Dac, N. N., Carson, S. J., & Moore, W. L. (2013). The effects of positive and negative online customer reviews: Do brand strength and category maturity matter? *Journal of Marketing*, 77(6), 37–53.

- Homburg, C., Vomberg, A., & Muehlhaeuser, S. (2020). Design and governance of multichannel sales systems: Financial performance consequences in business-to-business markets. *Journal of Marketing Research*, 57(6), 1113–1134.
- Hox, J. J. (2010). *Multilevel analysis – techniques and applications*, 2nd edn. New York: Routledge.
- Jacobson, R., & Aaker, D. A. (1985). Is market share all that it's cracked up to be? *Journal of Marketing*, 49(4), 11–22.
- Kennedy, P. (2008). *A Guide to econometrics* (6th ed.). New Jersey: Wiley.
- Klingemann, W., Kim, J., & Füller, D. (2021). Willingness to pay. In C. Homburg, M. Klarmann, & A. Vomberg (Eds.), *Handbook of market research*. Cham: Springer.
- Kummer, M., & Schulte, P. (2019). When private information settles the bill: Money and privacy in Google's market for smartphone applications. *Management Science*, 65(8), 3470–3494.
- Lovett, M. J., Peres, R., & Shachar, R. (2013). On brands and word of mouth. *Journal of Marketing Research*, 50(4), 427–444.
- Luo, X., & Bhattacharya, C. B. (2006). Corporate social responsibility, customer satisfaction, and market value. *Journal of Marketing*, 70(4), 1–18.
- Manchanda, P., Packard, G., & Patabhiramaiah, A. (2015). Social dollars: The economic impact of customer participation in a firm-sponsored online customer community. *Marketing Science*, 34(3), 367–387.
- McAlister, L., Srinivasan, R., Jindal, N., & Cannella, A. A. (2016). Advertising effectiveness: The moderating effect of firm strategy. *Journal of Marketing Research*, 53(2), 207–224.
- Misangyi, V. F., Elms, H., Greckhamer, T., & Lepine, J. A. (2006). A new perspective on a fundamental debate: A multilevel approach to industry, corporate, and business unit effects. *Strategic Management Journal*, 27(6), 571–590.
- Mizik, N., & Jacobson, R. (2008). The financial value impact of perceptual brand attributes. *Journal of Marketing Research*, 45(1), 15–32.
- Mizik, N., & Pavlov, E. (2021). Assessing the financial impact of brand equity with short time-series data. In C. Homburg, M. Klarmann, & A. Vomberg (Eds.), *Handbook of market research*. Cham: Springer.
- Naik, P. A., & Raman, K. (2003). Understanding the impact of synergy in multimedia communications. *Journal of Marketing Research*, 40(4), 375–388.
- Nath, P., & Mahajan, V. (2008). Chief marketing officers: A study of their presence in firms' top management teams. *Journal of Marketing*, 72(1), 65–81.
- Petersen, M. A. (2009). Estimating standard errors in finance panel data sets: Comparing approaches. *The Review of Financial Studies*, 22(1), 435–480.
- Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using stata, volumes I and II: Multilevel and longitudinal modeling using stata*. Texas: Stata Press.
- Rao, V. R., Agarwal, M. K., & Dahlhoff, D. (2004). How is manifest branding strategy related to the intangible value of a corporation? *Journal of Marketing*, 68(4), 126–141.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models – Applications and data analysis methods*. California: Sage.
- Roodman, D. (2009). How to do xtabond2: An introduction to difference and system GMM in Stata. *The Stata Journal*, 9(1), 86–136.
- Skiera, B., Reiner, J., & Albers, S. (2021). Regression analysis. In C. Homburg, M. Klarmann, & A. Vomberg (Eds.), *Handbook of market research*. Cham: Springer.
- Sorescu, A. B., & Spanjol, J. (2008). Innovation's effect on firm value and risk: Insights from consumer packaged goods. *Journal of Marketing*, 72(2), 114–132.
- Srinivasan, S. (2021). Modeling marketing dynamics using vector autoregressive (VAR) models. In C. Homburg, M. Klarmann, & A. Vomberg (Eds.), *Handbook of market research*. Cham: Springer.
- Steenkamp, J. B. E., & Geyskens, I. (2014). Manufacturer and retailer strategies to impact store brand share: Global integration, local adaptation, and worldwide learning. *Marketing Science*, 33(1), 6–26.

- Valli, V., Stahl, F., & Feit, E. (2021). Field experiments. In C. Homburg, M. Klarmann, & A. Vomberg (Eds.), *Handbook of market research*. Cham: Springer.
- Verbeek, M. (2017). *A Guide to modern economics* (5th ed.). New Jersey: Wiley.
- Vomberg, A., & Klarmann, M. (2021). Crafting survey research: A systematic process for conducting survey research. In C. Homburg, M. Klarmann, & A. Vomberg (Eds.), *Handbook of market research*. Cham: Springer.
- Vomberg, A., Homburg, C., & Bornemann, T. (2015). Talented people and strong brands: The contribution of human capital and brand equity to firm value. *Strategic Management Journal*, 36(13), 2122–2131.
- Von Auer, L. (2013). *Ökonometrie – eine einföhrung*, 4th edition, Berlin Heidelberg: Springer.
- Wang, W., & Yildirim, G. (2021). Applied time-series analysis in marketing. In C. Homburg, M. Klarmann, & A. Vomberg (Eds.), *Handbook of market research*. Cham: Springer.
- Warren, N. L., & Sorescu, A. (2017). Interpreting the stock returns to new product announcements: How the past shapes investors' expectations of the future. *Journal of Marketing Research*, 54(5), 799–815.
- Wies, S., Hoffmann, A. O. I., Aspara, J., & Pennings, J. M. (2019). Can advertising investments counter the negative impact of shareholder complaints on firm value? *Journal of Marketing*, 83(4), 58–80.
- Wlömert, N., & Papies, D. (2019). International heterogeneity in the associations of new business models and broadband internet with music revenue and piracy. *International Journal of Research in Marketing*, 36(3), 400–419.
- Wolters, H. M., Schulze, C., & Gedenk, K. (2020). Referral reward size and new customer profitability. *Marketing Science*, 39(6), 1166–1180.
- Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. Massachusetts: MIT Press.
- Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach* (6th ed.). Massachusetts: Cengage Learning.