Problem Set 2

Saskia Jacob & Martin Klarmann

Winter Term 2025/26

Data For this problem set, please use the data from the previous problem set together with an additional dataset. For the first and second tasks, please use the data on the Fulton Fish Market that we already worked with last week. This dataset contains information on daily sales of whiting at the Fulton Fish Market in New York City, collected by Kathryn Graddy in April and May 1992, and includes detailed records on prices, quantities, and market characteristics.

For the third and fourth task, please download and use an additional dataset from the paper "Close a Store to Open a Pandora's Box? The Effects of Store Closure on Sales, Omnichannel Shopping, and Mobile App Usage" from Ye & Shankar (2024). The dataset "StoreData.csv" contains monthly panel data with 1,188 county-level observations covering an 18-month period, with months 1–6 representing the pre-closure phase and months 7–18 the post-closure phase. It distinguishes between counties affected by the closure and those serving as controls, enabling comparisons across treatment and non-treatment regions. The dataset includes detailed information on sales and returns (frequency, value, and quantity) across both online and offline channels. Due to confidentiality agreements, all numerical values have been modified by adding random noise and scaling factors, and location-specific identifiers have been removed. Thus, the data preserve the structure and relationships of the original study but not the exact numerical results. Detailed information about the individual variables can be found in the readme file.

Document This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

Relevant R packages For the tasks this week you will need the following packages in addition to the packages included in the standard installation: `vars, tseries, forecast, plm, lme4`.

Deliverables and Deadline

1. A Markdown Notebook with all the code and your argumentation. We may apply this code to the original dataset, so please make sure that all changes to the dataset are programmed in the notebook (and not done in Excel etc.).
2. A PDF document with your responses to all tasks (max. 15 pages). We recommend using your Markdown Notebook document to create your PDF file.

Results need to be uploaded to ILIAS in the relevant section on Friday, 5th December 2025, at 5pm.

Task 1 "Price and Demand" - 5 out of 25 points In this paper, Kathryn Graddy explicitly raises that there may be endogeneity concerns when it comes to analyzing the relationship between price and demand on the Fulton Fish Market (p. 215-217). She suggests to use weather data as an instrument.

Please explain in your own words how the relationship between price and demand could be affected by endogeneity. Use data and plausibility arguments to support your standpoint. Is weather data (rain, storms,...) a suitable instrument in this context? Please use data and plausibility arguments to discuss.

Then, please re-run the price demand function you identified as best in Task 2 of the first problem set using instrumental variables (try out more than storms). Conduct Hausman tests. Is there an endogeneity problem

in the data? Do you see other endogeneity problems not captured by your instrument?

For this task, please use the "Daily Fulton Fish Market Data". This task explicitly refers to chapter 2.1 of the lecture.

Task 2 "Clustered Data" - 5 out of 25 points The "Detailed Fulton Fish Market Data" dataset is a nested dataset. In particular, purchases are nested in customers. In this task, please redo the analysis of two of your moderated regression models from Task 3 in the first problem set using multilevel modeling. First, describe how you expect that explicitly modeling the hierarchical structure of the data will affect the results. Then, compare these expectations to the observed results and provide a discussion.

For this task, please use the "Detailed Fulton Fish Market Data". This task explicitly refers to chapter 2.2 of the lecture.

Task 3 "Sales Forecast and Relationship between Offline and Online" - 9 out of 25 points Accurate sales forecasts are essential for retailers to plan inventory, allocate resources, and adapt to changes in customer behavior. Events such as store closures can substantially alter sales patterns by redistributing demand across online and offline channels.

First, please select two counties from the dataset: one from the treatment group (affected by a store closure) and one from the control group (unaffected). Develop a suitable time series forecasting model for the total sales value (offline + online) of each selected county using the methods introduced in class. This will, among others, entail differencing for trends to make the time series stationary (if necessary), account for seasonality (if necessary), and selecting the appropriate number of lags. Please justify your decisions using suitable figures and tests as well as sufficient commentary. Finally, provide a forecast for total sales for each county for the three months following the last observation in the dataset.

Next, analyze the interdependencies between online and offline sales using a Vector Autoregression (VAR) model for both counties. Compare the dynamics between the treatment and control counties, discussing how the interaction between online and offline sales differs across the two contexts.

For this task, please use the "StoreData". This task explicitly refers to chapter 2.2 of the lecture.

Task 4 "Predicting Store Closure" - 6 out of 25 points Identifying early warning signals for potential store closures is crucial for retailers to take proactive measures, such as adjusting marketing efforts or real locating resources. Patterns in key metrics may indicate underlying performance issues that precede a closure in a county.

Please explore which factors could predict whether a county will experience a store closure. Therefore, please select and justify three variables that you believe might indicate an increased likelihood of closure (e.g., sales levels, online share, return rates, discounts, or changes in these indicators over time). Formulate clear hypotheses explaining why the chosen variables could serve as early warning signals for store closures.

Next, build a logistic regression model to predict the treatment status (store closure = 1, no closure = 0) based on your selected variables. Evaluate the model's performance (e.g., AIC, R^2) and interpret the estimated coefficients in light of your hypotheses. Finally, discuss whether your results support your initial expectations and what they imply about the characteristics of counties more likely to experience store closures.

For this task, please use the "StoreData". This task explicitly refers to chapter 1.4 of the lecture.