ECN431
Lab session 1
Empirical analysis starter

Aud 12
Friday, January 19th 2018

Before the Lab

- Make sure you have access to R on your laptop with necessary packages – you can find a helpful guide for this on the GitHub page for the course.
- Make sure you have the data file and the R notebook with examples and hints available locally. If you clone the git repository using the guide on the Github page of the course, these files will be included. If you cloned the repository prior to the files being posted, you can update your cloned repository on Github (see the Github page of the course for details).
- Try to read through the background information on the "case" below prior to the lab and start working on the exercises. This will greatly improve the usefulness of the lab session. Use the provided notebook *lab1_example.Rmd* to get started.

Purpose of this lab

To explore data on market and firm behavior with the purpose of understanding the important features, to reason about endogeneity issues, reason about what we can learn from observation, and be able to present your findings in a suitable form.

Learning goals

- Understand omitted variable bias
- Improve reasoning about endogeneity
- Understand indicator variables and regression
- Be able to produce informative tables and graphs

Practical information

The easiest way to work with the example R notebook is by opening it in RStudio. This file contains commands that you will need in order to answer the questions in these exercises, in addition to information on some new commands. Remember: If/when you are uncertain about the purpose, functionality and extensions to/options for some of the commands, just select the command (either where it's written in the code editor, or by writing it in the console) hit F1 in RStudio.

Exercises

Rossmann: Competition and promotional activities

Dirk Rossmann GmbH, commonly known as Rossmann, is Germany's second largest chain of drug stores after dm-drogerie markt (*dm*). Rossmann has over 3,500 outlets and 50,000 employees in several European countries, with a turnover of 8.4 billion EUR in 2016. Compared to pharmacies in Scandinavia, Rossmann (and it's competitiors) carry a larger variety of products, such as consumables, household products and other groceries. With its large operations, demand forecasting and logistics planning is important in the short run for cost minimization, while understanding the competitive situation and expansion choices through new outlets is important in the longer run for profit maximization.

In this exercise, you will use panel data on 1 115 Rossman stores in Germany from the file *rossmann.dta*. The data is recorded on a daily level from January 2013 to July 2015, and contains information on sales, number of customers, promotion campaigns, the competitive situation, records of holidays, and the store itself.

1. Rossmann has four different store concepts. The exact nature of them are not revealed in these data, though the variable *storetype* contains a classification. Tabulate the values of *storetype*. The values are stored as text (strings).

2. Regress sales on store concept (indicators for each). How would you interpret this?

3. Make a bar chart showing the share of time/days each store concept is open.

4. Regress sales on store concept and the indicator for being open. What is the interpretation of the coefficients? Explain the change in coefficients on store concepts, and try to show this using regressions and calculations.

5. Regress sales on store concept interacted with the indicator for being open. Interpret the coefficients, particularly noting the difference with the previous estimates.

6. Make a table of average sales per store concepts only for days the store is open, and explain how this relates to the coefficients you just estimated.

7. Regress sales on store concept only for days when a store is open. Explain the change in R^2 from the previous regression.[1] Under what circumstances will it be okay to limit the sample in this way? You can assume that you want to estimate a more complicated model.

[1] Once you're certain about your explanation, feel free to contemplate whether and when R^2 is useful.

8. Regress sales on whether the store has a competitor nearby (the variable *competition*), with and without controlling for whether the store is open. What do you make of the coefficients? Calculate what they imply in percentage changes. What could be a potential issue with controlling for stores being open in this regression? Taking the issue(s) you think could be relevant, try to investigate whether there is reason to be worried using the data.

9. Regress sales on distance to the closest competitor.[2] Make sense of the coefficient.

10. Create a graph or table showing the share of observations facing competition each year (the average of competition in each year). What does it tell us about variation in competition over time? Also create a graph or table showing average daily sales for each year (see if you can combine both figures/tables). What does this tell us about the regression of competition on sales?

11. Run the regression of sales on competition with store-fixed effects. Use the command `plm` from the package `plm`. How would you describe the variation in the competition variable used for estimating the coefficients in this regression compared to the regression without fixed effects?[3] Also add dummies for each year and compare the results. What could still be issues with interpreting the coefficient on competition as the (causal) effect of competition on sales?

12. Also run a regression of being open on competition with store-fixed effects. Would you say that you are more confident in these estimates informing us about how competition affects how often stores are open, compared to the ones without fixed effects?

13. Make a graph or table showing how common the two different promotional activities are for each of the store concepts. Promotions spanning the whole chain is indicated by *promo*, while store-specific promotion is indicated by *promo2*.

14. You want to regress sales on both type of promotions interacted with store type (can you explain why this could be a sensible thing to do?). Decide first if it is *necessary* to control for competition in this regression, and explain your choice. Run the regression both with and without store-fixed effects, and output the results side by side in a sensibly formatted table using `stargazer`. Interpret and try to find potential explanations for the differences in coefficient estimates between the two regressions.

15. Extra:

    (a) Generate a table of summary statistics for sales, customers, being open, promotion (both types), nearby competition, and distance to nearest competitor.

    (b) Calculate the number of stores who, during our sample period: a) Never face competition, b) Always face competition and c) Started facing competition

[2] Tip: Distance is measured in meters here. Generate a variable measuring it in km instead. Sometimes, coefficients are a bit difficult to read and sensibly interpret if they become "too" small or "too" big. If there is a sensible rescaling which alleviates this problem, you should do it.

[3] Hint: What are we partialling out here?

(c) Make a graph showing the number stores getting competition in each month (where competition goes to 1 from 0 in the previous month).

(d) For the stores who had a competitor enter nearby during our sample period, generate a counter for days centered on entry by competition (takes the value 0 on the day competition started, 1 the day after and so on, and -1 the day before and so on). Make a plot of average sales for each day relative to competitive entry among these stores in window of 90 days before to 90 days after. Interpret and comment. What do we learn about the regressions we ran from this plot?