Introduction to Text Mining and Natural Language

Problem Set 1

Guillem Mirabent Rubinat, Amber Walker, Alvaro Ortiz

22DM014 taught by Hannes Mueller

February 4, 2024

1 Identify a (future) event that makes a lot of people come to Barcelona. Think about music festivals, local festivities etc. (2 points)

We are choosing the "La Mercè" festivities because they are a big event in Barcelona which attracts a lot of tourism. This year, "La Mercè" doesn't have fully confirmed dates yet but as the main event is on the day of Saint Mercè (24th September), we can safely assume that it will take place around 22-25 September. In order to capture possible prearrivals in Barcelona and considering that on the 25th there's only minor events left, we are considering as treatment period the one comprised between 19-25 of September. As a control week we are using the period (of the same length) which is between 26 of September and 2 of October. Both periods are in September or in the very beginning of October so the possible differences because of seasonality are minimized (weather and summer vacation effects are the main ones we are worried about). Additionally, the control week is set after summer vacations in Catalonia and well into the normal working calendar. The Halloween effects stemming from parties and celebrations in Port-Aventura could be an issue but 2 Oct is still far away enough for it not to be a problem. As a result, regarding Barcelona both periods 19-25 Sept and 26 Sept - 2 Oct would be very similar if it wasn't for "La Mercè" festivities.

2 Think of the time periods to scrape and what second city to scrape for these same timer periods. Explain your choices in written. (2 points)

As a second city we are using Naples. It is a city which is quite comparable to Barcelona in terms of size. Barcelona is a little bigger but Naples is the most similar in size from cities in the Mediterranean with direct access to the beach and similar temperature and weather characteristics. Naples is a good comparable city as well because the only relevant festivity there in September is "San Gennaro", which is important for the local people but it is mainly religious so it is not as attractive for tourists. In comparison, "La Mercè" is a local festivity which spreads out to become a full festival, with huge stages for concerts on the beach and in general all over the city. The whole city of Barcelona becomes a festival.

As a result, we think that "La Mercè" can be a relevant tourist attraction to Barcelona for the treatment period while Naples does not have any relevant tourist attraction for any of the two periods.

We would also like to specify that we are focusing our search on rooms for 2 adults with no children. This is also a parameter that can be altered in Booking.com, so we decided to leave it as it is set by default in the webpage. It is the most common room setting in the hotels (i.e. rooms prepared for 2 adult people are the most common). As a result, we don't think this needs any tweaking from our part, but we do want to specify our choice. We maintained this choice for all the searches.

Figure 1: San Gennaro, Naples

Figure 2: La Mercè, Barcelona

3 Design a careful scraping pipeline that follows the advises seen in class and TAs. (5 points) The basic points to bear in mind are:

- Organize the data you need, format and structure to store it beforehand. Try to foresee how you will need to read in the data to answer your questions. If you want, you can include some few lines explaining your pipeline strategy at the beginning.

python code

- Codes should be as automated as possible. That is, you don't want to rely on human intervention to get your data.

python code

- Use only the packages we have seen in the course. Although firefox is recommended, you can also use chrome as your scraping browser.

python code

- Document your codes and make them robust and efficient.

python code

4 Scrape date, room price, hotel name and hotel description. (5 points)

python code

5 Write down three equations. One with just a "treatment period" dummy, one with just a "treatment city" dummy and then one that adds both of them with their interaction. This last regression gives you a difference-in-difference estimate of the effect of the event on prices. Explain which coefficient captures this treatment effect and why you need a second city for this. (3 points)

a) Treatment Period Dummy Regression: This model examines the impact of the event on rental prices, disregarding the city. We denote:

- P as the price.
- D_{event} as a dummy variable that equals 1 if the observation is during the event period, and 0 otherwise.

Equation:

$$P = \beta_0 + \beta_1 D_{\text{event}} + \epsilon$$

Here, β_1 shows us the average effect of the event period on rental prices. β_0 is the intercept and ϵ is the error term to allow for some noise in the regression.

b) Treatment City Dummy Regression: This model observes the discrepancies in rental prices between Barcelona and the control city, independent of the event. We use:

- $D_{\text{Barcelona}}$ as a dummy variable that equals 1 if the observation is from Barcelona, and 0 if from the control city.

Equation:

$$P = \alpha_0 + \alpha_1 D_{\text{Bcn}} + \epsilon$$

In this equation, α_1 measures the average price difference between Barcelona and Naples. α_0 is the intercept and ϵ is the error term to allow for some noise in the regression.

c) Combined Treatment Period and City Regression with Interaction: This is the key difference-in-differences regression. It includes both the treatment period and city dummies, plus their interaction.

- $D_{\text{event}} \times D_{\text{Barcelona}}$ is the interaction term.

Equation:

$$P = \gamma_0 + \gamma_1 D_{\text{event}} + \gamma_2 D_{\text{Barcelona}} + \gamma_3 (D_{\text{event}} \times D_{\text{Barcelona}}) + \epsilon$$

Here, γ_3 (the coefficient of the interaction term) captures the differential effect of the event on rental prices in Barcelona compared to the control city. This coefficient is the DiD estimator and shows the impact of the event after accounting for city-specific factors. γ_0 is the intercept and ϵ is the error term to allow for some noise in the regression.

d) **Why we need a Second City for Comparison:** A second city (control city), in this case Naples, is needed as a comparison to isolate the effect of the event in Barcelona. Naples is similar to Barcelona but not affected by the event. This comparison helps us distinguish the event's impact from other time-specific effects that could influence rental prices.

e) **Treatment Effect Capture:** In the DiD model, the interaction term's coefficient (γ_3) captures the treatment effect. It measures how much the event in Barcelona changed rental prices relative to the control city during the same period.

By analyzing these regressions, you can understand the isolated effect of the event in Barcelona on rental prices, controlling for city-specific characteristics and other time-specific factors that might affect the rental market.

6 Estimate all three regressions. Make a standard regression table with 4 columns (3 for your answer here and one more below). Make sure you check how these regressions look like usually. Always report all coefficients. Then carefully interpret them for each regression and the changes you see. (4 points)

	Dependent variable: price		
	Model 1	Model 2	Model 3
	(1)	(2)	(3)
EventPeriod_Barcelona			-30.229
			(245.661)
const	1557.679***	1067.373***	1084.634***
	(78.132)	(104.059)	(147.729)
is_Barna		637.371***	651.406***
		(122.721)	(173.357)
is_Mercè_time	-65.383		-34.308
	(111.615)		(208.279)
Observations	1153	1153	1153
R^2	0.000	0.023	0.023
Adjusted R^2	-0.001	0.022	0.021
Residual Std. Error	1894.606 (df=1151)	1873.067 (df=1151)	1874.475 (df=1149)
F Statistic	0.343 (df=1; 1151)	26.974*** (df=1; 1151)	9.069*** (df=3; 1149)
Note:			*p<0.1; **p<0.05; ***p<0.01

For model 1, the coefficient is -65.383, which we found as not statistically significant. This suggests that during the La Mercè event, the hotel prices are not affected by the event and generally stay the same. Thus, we were unable to find any evidence of the La Mercè event driving hotel prices up.

For model 2, the coefficient is 637.371, significant at the p-value of 0.01 level (***). This indicates that being in Barcelona (as opposed to Naples) increases rental prices by about 637 euros.

The interaction coefficient, 'EventPeriod_Barcelona', is -30.229, not statistically significant. It suggests that during the event period and being in Barcelona, the hotel prices are generally not affected and we see no evidence of the La Mercè event driving hotel prices up.

7 For each of the hotel descriptions do the following:

a) Extract at least two text features that can be useful controls in the regression. Think about the methods covered in class to transform text into numeric features and explain your decision. Show summary statistics for your feature for the different cities and time periods. (10 points)

* First we decided to extract three dummy variables from the different descriptions which were related to key aspects which are usually drivers of price:

  - "Free Cancellation"
  - "Breakfast Included"
  - "No Prepayment Needed"

  Those three "advantages" that some hotels offer are essentially additional services or flexibility options so they all might have a price of their own. As such, we extracted a dummy variable for each one of them.

* Second, we realized that a lot of hotels use the descriptions as a way to further expand on explaining their facilities and other perks that they offer. Additionally, the description also scraped a list of the facilities that the hotel selected to be shown right below the description on their individual web. As such, we decided to look at the most relevant "facilities" filters on Booking.com and count the number of times any of those appeared on the description of every hotel. For example, if an hotel mentioned that they have a "swimming pool" and offer "air conditioning" in the rooms, they would have a count of 2. From this count, we extract a numeric value which represents the level of "Facilities" of each hotel. This "level" is computed as follows:

$$Level_{\text{facilities}} = \frac{Count_{\text{facilities}}}{Length_{\text{description}}}$$

Where the count is the simple count of the number of facilities mentioned in the description (as explained above), the length of the description is not computed directly but rather what is counted is the number of "relevant" words. The way we computed this length was by applying lemmatization and stopword removal to the main text of the description. This way we are skipping words that don't really add to the meaning of the description. To make the column more easily understandable, we also standardized the results after normalizing.

b) Run a 4th regression and add it to your table. This should add your controls to your 3rd regression. Interpret the coefficient on the controls and the change you see on your treatment variable. (2 points)

Equation:

$$P = \gamma_0 + \gamma_1 D_{\text{event}} + \gamma_2 D_{\text{Barcelona}} + \gamma_3 (D_{\text{event}} \times D_{\text{Barcelona}}) + \gamma_4 X_1 + \gamma_5 X_2 + \gamma_6 X_3 + \gamma_7 X_4 + \epsilon$$

Where:

- X_1 is a dummy variable indicating whether the hotel offers free cancellation.
- X_2 is a dummy variable indicating whether the price includes breakfast.
- X_3 is a dummy variable indicating whether prepayment is required.
- X_4 is the 'level' of facilities included in the description.

	Dependent variable: price			
	Model 1	Model 2	Model 3	Model 4
	(1)	(2)	(3)	(4)
Breakfast_Included				562.416***
				(168.104)
EventPeriod_Barcelona			-30.229	-36.273
			(245.661)	(245.825)
Facilities				43.162
				(56.667)
Free_Cancellation				-137.777
				(155.723)
No_Prepayment				95.222
				(192.055)
const	1557.679***	1067.373***	1084.634***	846.442***
	(78.132)	(104.059)	(147.729)	(178.561)
is_Barna		637.371***	651.406***	876.852***
		(122.721)	(173.357)	(192.315)
is_Mercè_time	-65.383		-34.308	-27.407
	(111.615)		(208.279)	(208.244)
Observations	1153	1153	1153	1153
R^2	0.000	0.023	0.023	0.034
Adjusted R^2	-0.001	0.022	0.021	0.028
Residual Std. Error	1894.606 (df=1151)	1873.067 (df=1151)	1874.475 (df=1149)	1867.333 (df=1145)
F Statistic	0.343 (df=1; 1151)	26.974*** (df=1; 1151)	9.069*** (df=3; 1149)	5.746*** (df=7; 1145)
Note:				*p<0.1; **p<0.05; ***p<0.01

- Breakfast_Included: The coefficient is 562.416, highly significant (***), indicating that having breakfast included increases the rental price by about 562 euros.
- Free_Cancellation: We found this control variable not statistically significant due to the p-vaule being more than the threshold, suggesting it may not have a distinct impact on hotel prices.
- No_Prepayment: Same as with the free cancellation variable, we also found the no prepayment variable to not be statistically significant.
- facilities: The coefficient is 43.162 and not statistically significant.
- is_Barna: The impact of being in Barcelona remains significant and large (876.852).
- is_Mercè_time: The coefficient is small and also not significant in this model.
- The fourth model includes all variables. EventPeriod_Barcelona: The interaction is similar to Model 3 but slightly less in magnitude (-36.273).

Overall interpretation: We found that being in Barcelona as opposed to Naples significantly increases hotel prices. Including breakfast as a service in the hotel increases prices significantly. The La Mercè event seems to have no effect on prices in simpler models as well as more complex models with controls.

We found that other amenities like free cancellation and no prepayment do not show a significant effect in this model. The impact of the number of facilities is not statistically significant either, suggesting that it may not be a significant factor in determining hotel prices.

Some Notes:

- The R^2 values increase from Model 1 to Model 4, which tells us that adding more variables helps explain more variability in rental prices. That being said, the R^2 is still too small to be relevant enough.
- The significance of the coefficients can be interpreted based on p-values: *** for $p < 0.01$, ** for $p < 0.05$, and * for $p < 0.1$.

8 Imagine that you instead run the regression with hotel fixed effects (no need to run). Explain why the treatment effect will change and why your regression with controls should be closer to this regression. (2 points)

As we mentioned on the scraping notebook when presenting the concept of how we approached the web scraping part, we already selected only hotels in order to get as close as possible to this hotel fixed effects. In general, hotels, when looking at a period which is in 8 months, have many rooms of the same type available so they are much less likely to be sold out for one of the two periods we want to compare. It is much more probable that an apartment has one day already reserved during those 2 weeks we are scraping and, as a result, doesn't show in the search results for a given week while still showing up for the other.

That being said, as this measure is not perfect, the treatment effect would change because right now we still have some hotels that might be showing up for one week but not for the other. If we were to fix the hotel effects, we would essentially end up with the actual comparison of the prices of the same hotel instead of the sort of grouped comparison we are doing now. These hotels that show up only for one of the two weeks are introducing some kind of additional noise into our regression so eliminating them from the background would remove this noise, thus (probably) changing the treatment effect as a result.

With the controls, our 4th regression should be closer to the hotel fixed effects case because we are controlling for some of those hotel effects and fixing part of this noise. Our controls allow us to take into account how the prices are affected by the variation of the 4 different options we are considering (Free Cancellation (i), Breakfast Included (ii), No Prepayment Needed (iii) and the amount of Facilities (iv) offered by the hotel). Once we add those factors to the regression, they get specified and, as such, are no longer able to insert additional variation to the main regressor we are trying to study.