Winners and Losers from an Announced Excise Tax Hike: Tesla in Denmark
Online Appendix

January 5, 2018

1 Data appendix

This appendix, taken directly from (CITE) which uses the same data (CITE), explains in detail how our data was scraped, cleaned, and collected. We choose to include this discussion here rather than refer readers directly to the original thesis, because the original thesis may be difficult for some readers to access:

The data on used-car sales was derived from a web scraper that automatically saved the publicly accessible Bilbasen.dk in xml format. The files contained the html code of the Danish used-car websites search results for Tesla automobiles posted for sale between 2015-10-01 and 2016-05-05. The scraper was programmed to run every half an hour and it was implemented after Bloomberg had published news on the abolishment of the 180% tax exemption of electric cars in their September 29 article Teslas Hit by 180% Danish on Cars as Green Goals Ditched Bloomberg 2015.

1.1 HTML Parsing

The data parsing process was predominantly programmed in R. However, before the xml files were read into the R environment, a few cleaning operations were necessary to secure that the text files are correctly interpreted by the R commands. Firstly, 40 files were corrupt and contained no or erroneously coded data. These files have been removed from the dataset as they presented no usable information. Secondly, the html code in the xml files was formatted in a way that prevented the R package XML and xml2 to recognize it as an html nodeset. This was due to the appearance of double double-quote marks () instead of the normally used single double-quote marks (). This issue was circumvented by running a perl code in the Mac OS X Terminal environment. The code implemented a Find and Replace functionality on all xml files, replacing the unwanted double double-quote marks with single double-quote marks. After the necessary file cleaning procedures, the xmls were ready to be read into the R environment to extract all relevant information for further analysis. To successfully gain out all the relevant information contained in the xml files, the XML and xml2 packages were used within R. These two packages integrate the XPath language in R that helps navigate the user through the specific parent-children node-structure of the given html codes. With the help of the XML package, it was possible to search for specific html nodes with well-defined attributes. The variable information was contained in two forms: either as the value of a node or as the assigned value of an attribute of a node. With the XPath search commands, it was possible to extract both the nodes value or any of the necessary node attributes that contained the above variables. Since each element of a car sales advertisement was duplicated in the xml file, it was necessary to get rid of these duplicates before coercing the data from each individual file. To parse through the approximately 9000 files with an average of ca. 100 records each, a regular for loop was constructed in R that repeated the programmed operations on each xml file in the directory. There were 7 different vectors of information extracted from the xml files, one variable available for each record of advertisement.

(1) urls: these served as the unique identifies (IDs) of each separate car for sale (2) title: the title was essentially the model number (example: Tesla Model S P85D) (3) price: the listing price of the respective car for sale (4) years: the manufacturing year of the respective car for sale (5) kmeters: the odometer information available in the respective advertisement (6) region: the sales region indicated in the respective advertisement (7) descriptions: the description attached to the respective advertisement

In addition to the 7 variables listed, the filenames were added to each record to identify the exact datetime belonging to the advertisements. For this purpose, the filenames were conveniently formatted as yyyymmdd-hhmmss.xml from the source date of the file, which allowed a direct transformation to a datetime variable in R. The resulting 8 variables were coerced into a data frame object, where each row contained one record of advertisement from the parsed xml files.

1.2 Data Cleaning

After having parsed through the xml files, the program returns a total of c.a. 950.000 observations. Since the xml files were scraped every half hour from the source site, most of these observations are duplicates of previously occurring advertisements. The unique URL IDs allow the identification of duplicate advertisements.1 The different treatment of duplicate values results in three separate data tables that are used in different parts of the Summary of the Data and the Analysis chapters later. The three tables are constructed as follows:

(1) Leaving out all duplicate values based on the URL identifiers will result in having a data table with a list of all individual announcements. This can be taken either from the first or the last observation, providing a table with new announcements or delistings respectively.

(2) A second way to deal with duplicates is similar to the first method, however it sets slightly looser conditions for filtering out duplicate values by

allowing those advertisements that experienced a change in their price to remain in the sample as duplicates. This will allow us to retain more data points with useful information, that is the changing price over time of the same individual car.

(3) Similarly to the previous methodologies, we filter out duplicate values, however this time the data is first tabulated based on dates. The resulting data table will essentially contain each days list of advertisements, giving the opportunity to analyze daily statistics.

The next step in the data cleaning process was to detect and filter out the observations that cannot be interpreted in the analysis. Notably, prices with the Ring value and amounts under 100.000 DKK are discarded since these advertisements often indicate leasing offers, whereas we are interested solely in the direct sales data of Tesla automobiles. Finally, a filter is imposed on the title field (model number) to provide clarity in later interpretation of the analysis; the two Model Xs and three Tesla Roadsters are excluded, hence only Tesla Model S vehicles are kept in the final dataset.

The title of the observations provides useful information that are grouped into factor and binary variables. It was possible to identify the battery package (e.g. 60 kWh), whether the advertised car has front and rear motors (dual-motor) and whether the car is equipped with the performance package that is available for the high-end models and improves the cars acceleration and top speed. Furthermore, it was also possible to retain information on whether the car was equipped with an extra row of seats (sevenprs) and whether the car is one of the firstly released, signature edition models (signaturemodel). Even though there are other internal and external design and functionality upgrades, these five features represent the most significant factors of variation in the price of a Tesla Model S. Since it is possible to identify the first day and the last day of individual advertisements from the data. From this information, it is also possible to determine the number of days a car spent on the used-car market. For this purpose, a time-on-market variable is assigned to each record in the dataset.

As a last step in the data preparation procedures, the event variables were to be defined and assigned to each record in the dataset. Two event dates were selected for further analysis: the date of announcement (9th of October) and the date when the new legislation came into force (1st of January). Two approaches are used to create the event variables. First, the dates of the announcements are dummy coded; 0 for pre-event occurrences and 1 for post-event occurrences. Second, a week distance (weekdist) variable is constructed by taking the difference in weeks from the event date; negative for the weeks before the event and positive for the weeks after. Constructing the weekdist variable for the October 9th event is omitted due to availability of the data only after the 1st of October.

The final list of variables: dates; urls (unique id); title (char); price (num); year (factor); kmeters (num); region (factor); description (char); model_number (char); battery (factor); dualmotor (binary); sevenprs (binary); performancepackage (binary); timeonmarket (num); jan1_dummy (binary); oct9_dummy (binary), weekdist (numeric).

1.3 Additional Data Sources

As an additional data source to the used-car dataset, new car registration data was requested from the Danish Car Importers Association (De Danske Bilimportrer). This data included all newly registered personal vehicles in Denmark, including registration numbers on separate models, brands and their fuel type. The dataset spanned from 2015 to the early months of 2017, including all relevant periods of the analyzed legislative change. The available granularity was at a monthly level and the data did not include the prices of the registered vehicles. For prices and registration tax amounts on different Tesla models, we have relied on the quoted prices from the official website of Tesla Denmark Teslamotors.com 2017.

2 The luxury electric vehicle tax

For luxury electric cars, there was to be be no phase in, and the full registration tax was to be charged starting January 1, 2016. Specifically, the full tax would be assessed on automobiles priced higher than 800,000 DKK. This includes higher-end versions of Tesla Model S. Under that bill, the Table 1 shows how the after tax price of a new luxury Tesla Model S would have changed relative to other electric cars sold in Denmark. The price of a new luxury Tesla shoots up 74% in 2016. The European Commisssion, however, deemed the Danish tax

Car Type	2015	2016	2017	2018	2019	2020
Tesla S P85D	875	1522	1585	1702	1812	1807
VW e-Golf	287	301	318	344	351	334
Nissan Leaf	274	295	318	351	365	352

Table 1: Projected after-tax prices of electric cars in Denmark in thousands of DKK as of 9 October 2015

bill in violation of European Union competition law on 17 December, 2015 (DR article cited in thesis). The differential taxation of luxury electric cars like the Tesla Model S created an implicit subsidy for the smaller electric cars produced, for example, in Germany. This portion of the bill was officially omitted by the Danish parliament on the 27th of December.