Chapter 1

Data sources and main features

1.1 Bidding Data

1.1.1 Source and cleaning steps

Our main dataset is a set of firm proposals submitted in public auctions procured in Chile by government units between 2010 and 2021. Each observation corresponds to a proposal submitted by a firm to an auction for a specific contract. Each observation includes project characterization variables, auction characterization variables, and firm characterization variables.[1].

Raw data for public purchases developed via Mercado Publico (the digital platform where most of public procurements procesess are developed) is publicly available in the Open Data Portal of the Directorate of Public Purchases[2]. As mentioned in the Institutional Context section, most government units are mandated by the law to develop their procurement process via Mercado Público. Additionally, units who do not use the portal to purchase goods are still mandated by law to publish a basic set of information to the database. Given the law requirements for firms to submit purchasing data to the platform, we expect this dataset to include all purchases made by government units in the construction type save for some exceptions mentioned at

[1]Note that first two sets of variables are the same within bids for the same contract
[2]https://datosabiertos.chilecompra.cl/, last visited, july 2021

the end of the section.

The Open Data Platform has available data on public purchases in .csv files covering one year-month of purchases each. The .csv files were downloaded and merged together to form an initial raw dataset of around 10,000,000 observations, which includes a much wider array of product categories purchased by the government than just construction projects.

The dataset has the following sets of relevant variables:

- Project characterization variables: auctions' date, geographic region, the product category, legal size classification, procuring government unit and government estimate of cost.
- Proposal characterization variables: unique tax identifier of the submitting firm, amount of the proposal (bid), amount of units awarded, acceptance status of the bid (accepted/rejected), and awarding status of the bid (winner/ loser).

The firm unique identifier can be of two types depending on the firm. For firms constituted as legal entities separate from final taxpayers(i.e. individuals), the unique identifier is the unique tax number given by the internal tax bureau. For firms identified with a final taxpayer, the unique identifier is the personal unique ID (RUN) that uniquely identifies every person in Chile. Therefore, the variable that allows us to follow entities across the years and contracts has very little noise in it and is subject to almost no errors. Government Unit's IDs are also tax identifiers, which save for extraordinary circumstances also should stay the same over the years.

The acceptance/awarded bid variables indicate whether the bid did pass the first screening for formal requirements. The awarding status variable indicates whether the proposal won the contract under auction.

We now detail the filtering steps employed to produce our analysis sample. First, we keep only only projects with "Construction Projects and Services" or "WORKS" in one of the product category standardized classification ("RUBRO2"). The vast majority (almost 90%) of our data comes from observations in the first category. The second category begun being employed in 2017 to identify auctions from the Ministry

of Public Works. These filtering steps render around 270,000 observations.

We filter projects where more than one item is awarded to a single contractor or any contractor offered more than one item. This helps to filter out materials-only contracts and keep actual construction works.

We also drop contracts with a government estimate of less than 20,000,000 CLP or where the maximum bid is less than 15,000,000 CLP (if there is no government estimate we do not take the first condition into account). This step aims to exclude excessively simple projects, like small repairs, which do not entail either relevant subject-matter or public-specific domain expertise.

Finally, we observe firms in the dataset with more than one offer for the same contract, since contractors are allowed to modify their proposals until the end of the auction. We keep only the last proposal by the same contractor in the same project when we have proposals from different dates. If we have multiple proposals with the same submission date and we cannot distinguish which was the last submitted one, we prioritize the selection of the one that won (because that would mean it was the definitive one), the one that was accepted (by the same rationale) and, as last resort, we select one randomly[3].

We end up with 163,626 observations, submitted for 49,449 unique projects.

We expect our dataset to miss contracts related to national security, for example, the construction of naval bases. However, we still see some contracts procured by the military, which probably do not have national security connotations. Second, we do not have complete data for the Ministry of Public Works. This Ministry is exempt from the specific rules of law related to public purchases since it has its own set of regulations governing procurement of projects in road, airport, and other types of projects. Although the law mandates that even in this case the Ministry should publish basic information to the digital platform mentioned in the previous section, in practice we observe that the information is only partial, especially before 2017.

[3]This filtering step, since we already know if the bid was accepted with certainty and wether the project was awarded with certainty, only introduces possible error in the bid amount of a proposal. In the part of the analysis where the bid amount is relevant, we make a correction to exclude these noisy bids from the analysis.

1.1.2 Description of Buyers, Sellers and Projects

This section describes some relevant features of our main dataset after cleaning steps. Here we characterize the main sample which is employed in the next chapter, altough different analysis in the chapters perform can include small adjustments which are detailed in time.

First we characterize the buyers. Table 1.1 shows relevant statistics regarding government units. We have 928 unique government organisms developing on average 53 auctions each across the 12 year period. Note that the average number of years in the sample for a government unit is six, which means good time coverage. We characterize the types of governmnet bodies in the sample by matching category strings to the unit's name. We find the distribution of units in table 1.2. It can be seen that municipalities make the most of the projects in the sample, followed by ministries. We observe some universities owned by state as buyers as well.

Table 1.1: Government Bodies Descriptive Statistics

name	N	Complete Cases	mean	std	max	min
Number of Auctions Performed	928	1	53.3	118	2780	1
Total Firms Submitting Proposals	928	1	61.5	80.5	1170	1
Years in the dataset	928	1	6.01	4.24	12	1
Average Firms per Auction	928	1	4.36	15.3	466	1

Table 1.2: Types of Government Bodies Developing contracts

Type of Government Body	Number of Contracts Performed	Percentage	Cumulative Percentage
Municipality	36359	74%	74%
Ministry	8382	16%	90%
Other	1550	4%	94%
Child School Board	1335	2%	96%
University	892	2%	98%
Police, Investigations	728	2%	100%
Regional Government	135	0%	100%
Army,Navy	100	0%	100%

Next, we describe sellers (firms) and their bids. Table 1.3 shows descriptive statistics for bids and firms found in the analysis sample. The average firm bids in ten projects and wins approximately two, which gives a mean winning share of areound

.21. This shows that winning projects is not easy for firms in the market. Note that the standar deviation is high (.29) which speaks about heterogeneity in the market.

Table 1.3: Sample Descriptive Statistics

name	N	Complete Cases	mean	std	max	min
Bid (all)	145000	1	1.44e+10	5.26e+12	2e+15	0.5
Winning Bid	36200	1	2.62e+08	2.47e+09	2.47e+11	1
Difference between 1st bid and 2nd (%)	36200	0.697	0.084	0.151	1	0
Number of Bidders per Contract	36200	1	3.26	2.32	23	1
Year	36200	1	2016	3.1	2021	2010
Offers made by Firm	15000	1	9.66	27.4	1930	1
Win prob. by Firm	15000	1	0.209	0.293	1	0
Offers won by Firm	15000	1	2.42	5.93	140	0

The time dimension is essential in the current investigation since we follow firms across time for our main research question to compute experience and outcomes. Table 1.4 displays the number of observations, unique firms and unique contracts for each year of the sample, along with key variables. As expected, contracts have increased over the years, but our sample seems does not miss years in the data.

Table 1.4: Number of firms and contract per sample year

Variable	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Number of firms	2501	2765	3038	3025	3155	3978	3677	3583	3436	3530	3700	1828
Number of Auctions	2372	3661	4710	3759	4110	5352	4640	4360	4280	4922	5313	2002
Firms per Auctions	7.27	5.67	4.87	4.26	4.43	17.5	4.61	5.34	4.94	4.94	5.77	4.23

1.2 Awarding Criteria Data

The dataset presented in the previous section does not contain variables related to the awarding criteria employed by government units to score proposals. The main research question requires that we are able to tell when was experience an explicit factor in the awarding decision, because in these cases it is trivially true that past experience helps to increase the probability to win a contract. We obtain information about this criteria by employing the Mercado Publico API, and the awarding minute.

We query the official API of Mercado Público with each contract in our main dataset. The API allows to extract the URL of the awarding minute of the project.

The URL is employed to download the full awarding minute in html format, which is then parsed to extract the awarding criteria. Fortunately, the format of this awarding criteria is almost always the same across minutes (see an example in the Appendix).

Almost 89% of the sample contract ID's are matched sucessfully to a URL and 85% are matched successfully to their corresponding criteria. Altough failing to match a contract with its awarding criteria does not make us drop it from the analysis sample, it will impact the set of contracts employed for outcome computation. The final criteria dataset contains three variables: the unique identifier of the contract, the text of the criteria employed, and its weight, and is later merged to the main dataset.

We create two indicator variables by contract for the presence of price and experience criteria and two variables for the corresponding weights. We relate individual items to these criteria by matching strings (e.g. "exp" for experience-related items) since the field is non-standarized text. Figure 1-1 displays the proportion of projects that include price and experience with positive weight and the histogram of weights. The NA cases are the contracts for which we could not find a match. Around 60% of contracts do consider experience, 12% is missing, so for outcome computation we will employ around 30% of our dataset.

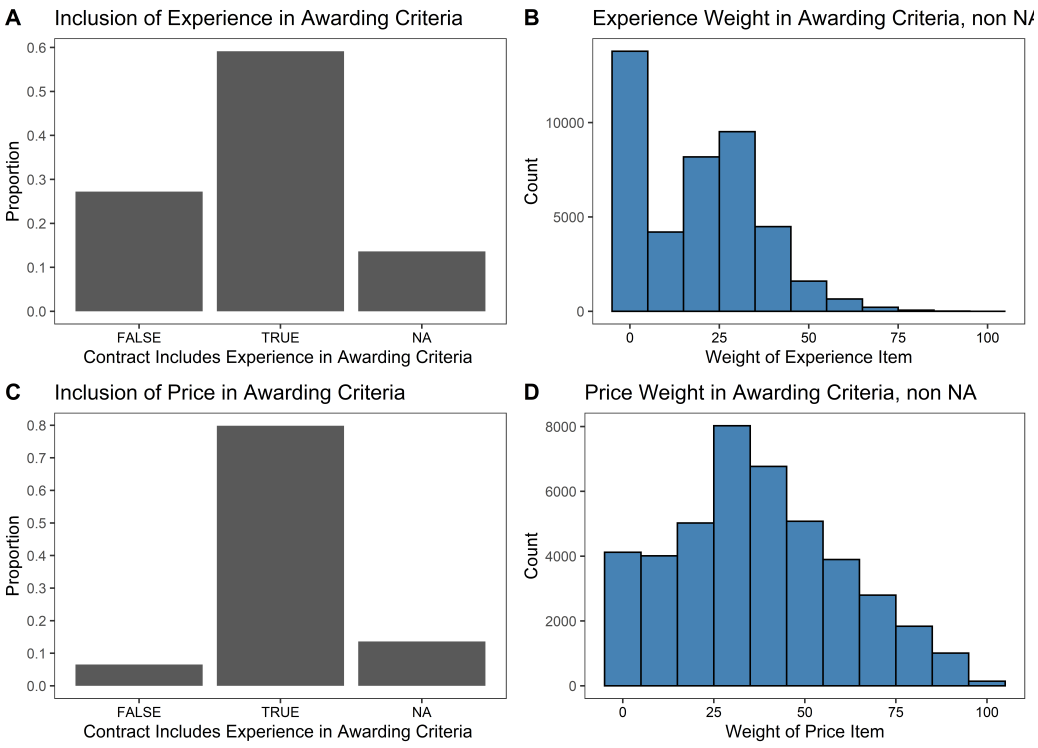
A
Inclusion of Experience in Awarding Criteria
Proportion
0.0
0.1
0.2
0.3
0.4
0.5
0.6
FALSE
TRUE
NA
Contract Includes Experience in Awarding Criteria
B
Experience Weight in Awarding Criteria, non NA
Count
0
5000
10000
0
25
50
75
100
Weight of Experience Item
C
Inclusion of Price in Awarding Criteria
Proportion
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
FALSE
TRUE
NA
Contract Includes Experience in Awarding Criteria
D
Price Weight in Awarding Criteria, non NA
Count
0
2000
4000
6000
8000
0
25
50
75
100
Weight of Price Item

Figure 1-1