Does host gender affect the prices of Airbnb listings? – A double machine learning approach

By

Benedek Tóth

Submitted to

Central European University

Department of Economics and Business

In partial fulfillment of the requirements for the degree of Master of Arts in Economics

Supervisor: Professor László Mátyás

Budapest, Hungary

2020

Abstract

This thesis investigates how the host's gender affects the price of listings on Airbnb. For this purpose, the double machine learning method is used to account for potentially nonlinear relationships between price, gender and other listing attributes. A standardized analysis is presented for 61 different locations, and the results are examined both together and individually. Gender effects are estimated for female hosts and couples in comparison to male hosts, and the relationships between these effects, host professionality and guest-host cohabitation in the listed apartment are also analyzed. These steps are necessary to disentangle different potential sources of gender effects, arising from either gendered pricing behavior or gender-sensitive demand. The thesis does not find evidence for the general presence of gender effects. Several individual coefficients in individual locations are significant but correcting for multiple comparisons invalidates this in the majority of cases. Only the effect of a professional, cohabiting hosting couple has a mean statistically different from zero (positive 3%) in the entire sample of results; this is also the group most likely to experience a positive gender effect based on the hypotheses presented in the thesis. The thesis also compares double machine learning estimates to linear regression and lasso coefficients and finds that the former is generally closer to zero and have a smaller variance.

Acknowledgements

I am thankful to Professor László Mátyás, my supervisor, for his comments and suggestions, but also for his patience and flexibility. I owe gratitude to Thomas Rooney for the innumerable improvements I made on the text with his help. My years preparing for and studying at CEU would have been much more grueling and less successful without Bea. The same is true for the students of Rajk College for Advanced Studies, who served as a constant inspiration. During writing this thesis, I always enjoyed Alina's support, which was a great help in completing this paper.

Table of contents

List of figures

List of tables

Introduction

Airbnb, as an online platform for bilateral arrangements, provides a unique environment to study real economic decisions. It connects numerous individuals looking for and renting out apartments and a high number of individual transactions are arranged through the platform. Due to the nature of its operations, it also creates a lot of data to analyze the outcomes of these transactions. As both the buyers and the sellers are (at least generally) individuals instead of professional organizations, outcomes potentially reflect personal attitudes irrelevant in deals including impersonal corporations.

There are notable studies taking advantage of this setting to analyze these attitudes. Edelman and Luca (2014), for example, find evidence for racial discrimination against African American hosts on the New York Airbnb market and Kakar et al. (2018) present similar results for San Francisco. Wang et al. (2015) study discrimination against Asian Americans in an analogous setting. Ert et al. (2016) study the impact of host photo availability on prices and find that hosts who appear more trustworthy charge higher prices on the platform. These are factors that are not present or at least not as clearly present in most market interactions between individuals and firms. Airbnb also presents an opportunity to observe "unprofessional" hosts making pricing and other entrepreneurial decisions. (Professionality, however, which itself is a vague term, probably greatly varies between hosts.) Oskam et al. in their 2018 paper, for example, examine the prevalence and effects of dynamic pricing on the Amsterdam Airbnb market.

This study examines the effect of host gender on Airbnb prices. This topic is covered less extensively in the literature than the effect of race. The aforementioned Edelman and Luca and Wang et al. papers do not report gender-specific effects, but Marchenko (2019) focuses on gender in addition to race in her study of seven US cities – although she does not find evidence for gender effects on prices. Unlike these studies, my analysis is not focused on the United States: I examine 61 different locations (mostly, but not exclusively cities), typically popular tourist destinations, from all over the globe. The aforementioned papers attribute the effect of race to discrimination, but potential gender effects could have various other potential driving forces as well. I hypothesize two effects I aim to measure, to which I refer to as a pricing-side and a demand-side effect.

The pricing-side effect arises from potential behavioral differences between female and male hosts: my hypothesis is that female hosts price their identical offerings less competitively on average, resulting in a negative effect from their perspective. (Such an effect would of course not be easily distinguishable from a similar negative effect that is the results of discrimination on the side of buyers and a corresponding pricing response on the side of hosts.) The demand-side effect is the consequence of hosts potentially having preferences over who to rent an apartment from. This would be driven by buyers' perceptions about trustworthiness and even more importantly safety, as Airbnb transactions often include meeting the hosts or even sharing the apartment with them. My hypothesis is that this results in a positive effect from the perspective of female hosts.

My hypothesis about the pricing-side effect is motivated by studies in different contexts showing that women behave less competitively than men in the situation in question. Thomas Buser and his various co-authors find evidence for such differences in an educational and career-choice setting (Busher et al., 2014, Busher et al., 2017 and Busher and Yuan, 2019), while Walter et al. (1998) analyze 62 studies about gender differences in competitiveness in bargaining situations and find that women behaved slightly less competitively in those. Heckman et al. (2009) presents experimental evidence that women are more risk-averse than men. Such effects of course might be highly context- or culture specific, and whether they apply to the Airbnb price setting example is not trivial – but pricing one's apartment reflects both competitiveness and risk aversion in this specific setting therefore analogous effects are plausible. Setting a higher price reflects competitiveness in the competition against other offerings as it implies a higher perceived value of the property and less risk-aversion, as a higher price results in higher chance of securing no bookings in a period all else held equal. It is important to note once again that similar effects could arise as a result of discrimination as well, and even behavioral differences might be rational reactions to buyer-side discrimination.

My hypothesis about the demand-side effect is based on buyers' perceptions about safety. Airbnb transactions generally require unsupervised meetings between the seller and the buyer and often even include them sharing an apartment. This makes some level of trust between buyers and sellers essential including the expectation on the side of guests that they are not going to be harmed or harassed during their stay. The various discussions of recent years about men abusing situations of economic dependence or interdependence and about women's sense of safety underline the importance of these factors. My hypothesis is that buyers – especially, but not exclusively women

– take them into account when choosing an Airbnb, and this produces a positive effect on the price of offerings with a female host.

The potential presence of effects with opposing signs presents a challenge: they are not directly observable separately and even if both effects exist their sum might be indistinguishable from zero. I propose three methods to separate these effects. The first is to compare offerings where the host(s) and guest(s) share an apartment (cohabit) with those where they do not. Safety concerns should be more important in the former case and under the assumption that the resulting demand-side effects are only present in this case, they are separable from pricing-side effects. The second approach is to compare professional and unprofessional hosts (as measured by the number of listings rented out by a host) under the assumption that the pricing-side effect is only present for unprofessional hosts as professional ones price effectively. (To assess the plausibility of this assumption I will also examine the effect of professionality on prices.) The third approach is to compare female and male hosts to couples, who are also present as hosts in considerable numbers, and assume that demand-side effects are present while pricing-side effects are absent for couples. Under any of these three assumptions the two effects can be separated – and estimates about the effects can also be used to assess how likely and compatible the assumptions are. (Obviously, intermediate scenarios where both effects are present but different for all groups are also possible.)

The aforementioned studies focusing on the effect of race on Airbnb prices generally use linear regressions with a number of control variables to produce coefficient estimates. This presents a serious potential problem if the underlying relationships between prices and price determinants are not linear, as the OLS estimates could be heavily biased in that case. This problem is not discussed extensively in the papers, but I aim to address it in my study. I use gradient boosting trees, a popular machine learning model, to estimate gender effects, as those are able to fit arbitrary patterns without a need for specifying any functional forms ex ante. To produce valid estimates using these models I follow the double/debiased machine learning method (DML), as presented by Chernozhukov et al. in their 2016 paper, “Double/Debiased Machine Learning for Treatment and Causal Parameters”.

This method relies on fitting predictive models of any kind on separate partitions of the data in question to produce one model predicting outcomes (price) and one model predicting treatment (gender). The resulting prediction errors are then regressed to produce the causal estimates. This

connects my study to several papers from various fields that utilize this relatively new method to estimate causal effects in high-dimensional and potentially highly complex settings. These include Knaus, 2018, where the author examines the effect of musical practice on cognitive skills; Daisuke, 2019 who studies the relationship between supply chain network structures and firm performance; and Yang et al. 2020, who study the "Big N" effect on audit quality. At the end of the paper, I present a brief comparison of double machine learning and OLS results, but I focus on estimating gender effects instead of evaluating the DML method.

In the remainder of the paper, I outline the data and data preprocessing steps used in this study, and the methodology of the analysis. Then I briefly summarize the results of the analysis for a single city for illustrative purposes, after which I discuss the entire set of results. I look at the OLS, Lasso and DML coefficient estimates both on an individual basis and as a realization of the sampling distribution of a single common coefficient. Finally, I compare the OLS and DML estimates and draw conclusions based on my research.

Data

In this paper, I analyze 61 separate but analogously structured cross-sectional data sets, each corresponding to a specific location. Every dataset contains a snapshot of the Airbnb offerings at the given location, the unit of observations being the properties listed on the site. For every observation, the data includes the price of the offering and many of its other relevant attributes, including ones describing the host(s) of the apartment. A host can have multiple corresponding offerings and therefore observations in a data set.

The data used in this analysis is provided by "Inside Airbnb", an open source project. This project is independent of the Airbnb company and aims to make the former's effects more transparent and understandable. They periodically provide information about all listings in a number of different cities and areas worldwide. This data is legally scraped from the Airbnb website and is available to download from Inside Airbnb's own webpage (Inside Airbnb, 2020a). It provides three tables for every location and date, which describe listing, calendar (the occupancy of properties) and reviews data.

I downloaded the listings table for every available location on February 7, 2020. This means the data I use was collected in the second half of January 2020, according to the Inside Airbnb site. (These tables are still available there, under the label of "archived data".) Prices could very well change systematically throughout the year, depending on the location, but there is no clear channel through which these potential seasonal changes would influence the effects I analyze in this study. It is also important to note that the data is from before the coronavirus outbreak was recognized as a global pandemic (World Health Organization, 2020). This pandemic clearly disrupted the Airbnb market, but my analysis considers a setting that is not yet affected by the virus.

The downloaded tables contain information on all listings that were available at the given location at the time of scraping. This includes the prices of the listings and several of their relevant properties. It is not trivial what I mean by prices in this context: the prices used in this analysis are nightly rates set by the host in the local currency. This includes neither additional fees (like a cleaning fee) and taxes, nor any possible discounts (for example one applied for a longer stay) (Airbnb, 2020). Some properties are provided for all locations, while others are only available for some of them.

Of these properties, in addition to price, I used the following in my analysis (generally as factors potentially determining the price, so either as controls, my variables of interest or as predictor variables): property type (apartment, house, loft, etc.); room type (entire apartment, private room, shared room, or hotel); number of guests accommodated, number of bedrooms, number of bathrooms, number of beds, bed type (real bed, couch, etc.); amenities listed (this can include anything from an air conditioner to a pool), the number of guests included in the price, the number of reviews, the mean review score rating, the number of reviews per month, the neighborhood of the listings, its latitude and longitude coordinates, whether the identity of the host is verified, the name of the host, and the number of Airbnb listings belonging to the host (calculated by Inside Airbnb).

An important missing variable here is the area of the listing, which is included in the tables but is not specified in the overwhelming majority of cases. For this reason, I do not use it in my analysis. Location data is somewhat imprecise, because Airbnb uses anonymization techniques to protect hosts. Therefore, the provided location is somewhere between 0-150 meters from the actual one (Inside Airbnb, 2020b). Effects of location differences on this scale would be completely idiosyncratic, therefore this imprecision presents no challenges in this context.

For some locations, the quality of the data is not sufficient for this analysis, so these locations have been dropped from my analysis. Overall 61 locations were suitable for the data preprocessing method I used (which is to be described later) therefore only those are included in the analysis presented here.[1] Neighborhood data is missing for several cities but was nonetheless included where it was available. This information is not directly provided by Airbnb, as it is added by Inside Airbnb based on the location of the listings.

Overall, the 43 following cities, metropolitan areas or other locations are included that do have neighborhood data available: Athens, Austin, Berlin, Bordeaux, Copenhagen, Denver, Dublin, Edinburgh, Florence, Hawaii, Hong Kong, Ireland outside Dublin, Istanbul, Lisbon, London, Los Angeles, Lyon, Madrid, Melbourne, Mexico City, Milan, Montreal, Munich, New Orleans, New York, Oakland, Oslo, Paris, Portland, Prague, Rio de Janeiro, Rome, San José, Seattle, Sevilla, Stockholm, Sydney, Taipei, Tokyo, Toronto, Venice, Vienna and Washington. In addition, the 18

[1] I note here that although not all the locations in question are cities, I often refer to them as such throughout the text – in these cases, I still mean locations in general.

following cities without neighborhood data are also included: Belize, Bergamo, Bologna, Geneva, Ghent, Girona, Jersey City, Málaga, Mallorca, Manchester, Menorca, Naples, New Brunswick (Canada), Northern Rivers (New South Wales, Australia), Ottawa, Porto, Providence (Rhode Island), and Sicily.

Data preparation was done separately for each city, according to the process described in the following paragraphs. The names of the hosts (which are usually only first names or even nicknames) are used to determine their gender, using the gender-guesser python package (Perez & Elmas, 2016). Names that are labeled as female or mostly female by the package are considered female, while names labeled male or mostly male as males. There are also several hosting couples (with names like Aaron and Alice, for example) present in the data.

Couples are identified as hosts with multi-word names, containing both a female and a male name. This means that by couples I only mean female-male pairs of hosts in this analysis. Many of the hypothetical effects could be present for same sex couples as well, but I aim to "use" couples to uncouple different gender-specific effects, and for that I need to look at heterosexual couples. Couples are quite rare among hosts, so it might not be possible to confidently identify effects for both female-male, male-male and female-female couples due to their low number in the data. Hosts with unknown gender (as determined by the algorithm) were dropped from the analysis. Generally, these observations do not seem to be systematically different from observations for which gender could be identified therefore I do not expect this to influence my results.

Non-numeric variables like property, room, and bed type are one-hot encoded as dummies. For regression analysis, the dummies representing the most common value of each variable are dropped and used as a "baseline case". The "amenities" variable is a list of several features for each apartment. Common features like TV, WiFi or even toilet paper have a unified naming convention on the Airbnb site, but other, more peculiar amenities also appear in the data. I generated a dummy variable for every amenity present in a given city; naming conventions are important here, as they ensure that there is only one single variable indicating the presence of an amenity. I also generated one additional variable: the overall number of distinct amenities by listing.

Some amenities or property types are very rare – to avoid overfitting on those with machine learning models, I dropped every dummy variable with less than thirty positive values. This choice

of a threshold number is admittedly arbitrary but is not expected to substantially influence the results of the analysis.

Possible room types include a shared room and a private room, both of which means that the guest has to share the apartment with the host(s) – these two dummies are combined into a single one, indicating the guest living together with the host. In regression analysis this case is use as the baseline and the variable is dropped. (I will occasionally refer to this variable or attribute of the host(s) and guest(s) living together as cohabitation.) The "number of listings by host" variable is also transformed into a dummy, simply indicating whether the host has multiple listings, which is to be used as a proxy for professional renting activities.

Prices are converted to their natural logarithm (on the one hand to obtain results easily comparable across currencies, and on the other hand because gender pricing effects, if present, are most likely multiplicative), and prices above a threshold are dropped: this threshold is different for every city – always the 99th percentile of prices for the given location. The reason behind this is the occasional presence of a few very highly priced offerings – these should probably not be considered as being on the same market with the other offerings (and compete with luxury hotels instead), but this difference might not be clear from the available variables. I removed these observations to avoid overfitting on their idiosyncratic properties. Observations with any of the variables missing (other than the exceptions discussed previously) were also dropped.

The steps described above were applied to every table (a table corresponds to one location) separately, but analogously. Those tables (and subsequently locations) for which this was not possible due to missing variables were dropped from the analysis

Methods

I performed the same analysis for all cities before evaluating the results as a whole. In this section I outline the methods, technical details and identification strategies. The goal of this analysis on the city level is to identify the effect of the host's gender on Airbnb prices, and how this effect depends on other factors – namely whether the host is "professional" and whether the guest and the host stay in the same apartment.

There are two challenges to this objective. On the one hand, it is important to control for factors other than gender affecting price to obtain more precise estimates – while knowing that the effect of these factors could take non-linear forms. On the other hand, the goal is to have unbiased estimates, and controls are necessary for that as well – it would not be prudent to assume that apartments rented out by hosts of different genders do not have systematically different attributes. Therefore, appropriately controlling for other attributes are necessary not only to obtain more precise estimates but also to obtain unbiased estimates. I used three methods for this purpose: simple linear regression, Lasso, and the "double/debiased machine learning method (Chernozhukov et al., 2016).

Linear regression

For the linear regression, a categorical dummy was dropped for all sets of categories: "house" for property type, "entire home" for room type and "real bed" for bed type. The variable indicating a male host was also dropped. Therefore, every effect is measured compared to these baselines.

A number of interaction variables were added to allow for heterogeneous effects. Every possible combination of the following variables was created as an interaction: "female host", "couple host", "professional host", "host and guest living together". This includes interactions of three variables, like the host being a couple, being professional and living with the guest. The rationale behind this is that the host gender could have a different effect if the guest needs to live with the host(s) and that professional hosts might respond to this in pricing differently than unprofessional ones. This potential effect would not be captured without these "triple" interactions. Seven interaction terms were added overall, as combinations where a host is both female and a couple are by definition impossible.

With these variables in place, I ran a linear regression with all variables included as controls. This serves as a baseline, an approach without any variable selection and with a simple implicit assumption about functional forms. There is no clear expectation as to how exactly price depends on property attributes or how those attributes are correlated with host gender. Therefore, it is not possible to confidently specify a functional form for these relationships ex ante, and a linear approximation seems to be a reasonable approach. It is possible, however, that this approximation leads to biased estimates (Chernozhukov et al., 2016), or even if it does not, a method accounting for general nonlinearities could possibly provide more precise estimates.

Lasso

As a second method, I use Lasso with OLS to perform automatic variable selection. This clearly does not solve the problems with potential nonlinearities discussed at the beginning of this section but serves another purpose. Running OLS for multiple cities, it should be expected that in some of those cases the estimated effects are significant at the 5% or even the 1% level even if the effects are uniformly zero across cities. This is solely due to the high sample size, viewing the estimates as a sample from the theoretical distribution of the estimated coefficient. (Of course, nothing guarantees that the true coefficients are the same or that the estimated coefficients have the same distribution across cities, but this leaves the core of the issue unchanged.) Lasso is used to provide a stricter and more clearly interpretable estimate on whether gender-specific variables have an effect on price, as it sets many of the coefficients to zero. The validity of standard errors obtained from a Lasso regression is ambiguous (Kyung et al., 2010); therefore I estimate a second linear model, only including the variables selected by the Lasso method previously, and report the standard errors from that model.

Double/Debiased Machine Learning

Third, I estimate the effects in question with the "double/debiased machine learning" method, proposed by Victor Chernozhukov et al. in the 2016 paper "Double machine learning for treatment and causal parameters" (Chernozhukov et al., 2016). This method is suitable for taking advantage of flexible machine learning models for estimation in a context where one is interested in the effects of a low dimensional parameter (the treatment) in the presence of high-dimensional nuisance parameters (Chernozhukov et al., 2016). The inference problem at hand fits this

description: I want to estimate the effect of gender (the treatment), possibly as a function of other low dimensional variables (two binary variables, “professionality” and “cohabitation) with a lot of potential confounders present that have an unknown functional relationship to price which should not be assumed to be linear.

The use of the term confounder in this setting is debatable: it is not clear whether one should interpret apartment properties as affecting the gender of the host (implying gender-specific decisions in choosing properties to purchase), or as being affected by the gender of the host (implying gender-specific decisions in furnishing a property), or both being affected by a third factor. It is clear, however, that the goal is to estimate the effect of the gender of the host independent of these other quantifiable factors. What I am interested in is how the host's gender affects her pricing behavior and the demand – basically the reservation price for an apartment – of the guests.

Using machine learning methods for coefficient estimation “naively” can result in heavily biased estimates, both because of the regularization bias inherent in ML models and because of the possibility of a flexible ML method overfitting on the idiosyncratic patterns in the data (Chernozhukov et al., 2016). The double ML approach solves these problems and results in an estimate with a number of desirable statistical properties. It produces point estimates that are approximately unbiased and normally distributed and are root-N consistent – meaning that the estimates concentrate in a $N^{-1/2}$-neighborhood of the true parameter value(s). (Chernozhukov et al., 2016) It also allows for calculating valid standard errors and confidence intervals.

In the following, I outline the practical steps of this estimation method – for more details, see the aforementioned Victor Chernozhukov et al. (2016) paper. The double ML estimation method requires the estimation of two predictive models: a model predicting the outcome using the controls or confounders, and a model predicting the treatment using the controls or confounders. These two models should not be fit on the same data; therefore, it is necessary to split the sample and estimate the two models on the two resulting disjunct datasets.

I use the generic term “predictive model” here, as the method allows for the use of any such model. This naturally includes popular machine learning methods, like neural nets and tree-based models (e.g. random forests and boosted trees). The two predictive models are then used to provide predictions for each observed unit, and the errors of these predictions are used to estimate the final

stage. The errors in outcomes are regressed on the errors in treatments (possibly allowing for this coefficient to depend on other variables) to obtain the coefficient estimates. This is summarized by the following equations (using a notation based on the Chernozhukov et al. (2016) paper):

1.) The assumptions on the data generating process, where Y is the outcome, theta is the treatment and X are the controls:

$$Y = D * \Theta(X) + g(X) + U \;\; E[U \mid X, D] = 0$$
$$D = m(X) + V \;\; E[V \mid X] = 0$$

2.) The predictive models and the resulting prediction errors:

$$\hat{Y} = \hat{g}(X) \; and \; \tilde{\mathrm{Y}} = \hat{Y} - \hat{g}(X)$$
$$\hat{D} = \hat{m}(X) \; and \; \tilde{\mathrm{D}} = \hat{D} - \hat{m}(X)$$

3.) The final stage regression, where $\Theta(X)$ has a specified form – the expectation used below describes the theoretical formula, while changing it to summation would describe the actual calculation, at least in the case of my analysis:

$$\hat{\Theta} = arg\,min\, E[(\tilde{\mathrm{Y}} - \tilde{\mathrm{D}} * \Theta(X))^2]$$

In the case of my analysis the outcome (Y) is the price of the Airbnb offering, the controls (X) are the various apartment properties discussed in the *data* section, and the treatment (D) is the gender of the host(s). $\Theta(X)$ is the effect of the host's gender on price. The assumed form of $\Theta(X)$ is: ($\alpha_0 + \alpha_1 professional + \alpha_2 cohabitation + \alpha_3 * professional\ AND\ cohabiting$) with different coefficients (alphas) for female hosts and couples, the two possible non-baseline values of D (the treatment) – in contrast to the baseline case to which all effects are compared (a male host).

For both predictive tasks, I used extreme gradient boosting, implemented by the XGBoost Python library (Chen & Guestrin, 2016 and xgboost developers, 2020). This is a version of the gradient boosting method (Friedman, 2002 and Mason et al., 2002), which stochastically generates and subsequently combines decision trees. The stochastic element of the algorithm is the repeated subsampling of both observations and features, meaning that at every step of the algorithm it only uses a subset of the entire input data. (For more details on the implementation see the sources cited above.) I chose this machine learning method based on my previous experiences on predictive problems and the impressive results achieved with XGBoost in public prediction competitions in the past years (Nielsen, 2016 and Fogg, 2016).

XGBoost fits a model to data based on several hyperparameters. I fitted each model (two per city) performing a hyperparameter grid search over a limited set of parameters to fit the best model possible (in terms of performance on the part of the input data that was not used to fit model). This grid over which the search was performed was chosen was based on experiments on a limited set of cities. Overall, the possible parameter values summarized in tables 1 and 2 (and consequently all of their combinations) were tried for each model.

Learning rate	0.03, 0.05
Maximal tree depth	10
Minimum child weight	1
Observation subsampling ratio	0.75
Variable subsampling ratio	0.75
Gamma (Minimum loss reduction required for a further partition on a leaf node)	0.1, 0.25

Table 1: Classification model hyperparameter options

Learning rate	0.015, 0.02, 0.03
Maximal tree depth	10, 15, 20
Minimum child weight	1
Observation subsampling ratio	0.75
Variable subsampling ratio	0.75
Gamma (Minimum loss reduction required for a further partition on a leaf node)	0.5
Number of estimators	1000

Table 2: Regression model hyperparameter options

The difference in the number of estimators is illustrative of the differences between the two predictive tasks. The output of classification is a categorical variable that can take a very limited

set of values, which increases the potential for overfitting if the number of estimators is high. A regression problem, on the other hand, requires more estimators in the model to fit continuous relationships with discrete steps. The gamma in this case, which specifies the minimum accuracy improvement necessary for further complicating the model, is higher (as based on empirical hyperparameter optimization) which helps counter the overfitting risks posed by the high number of estimators.

Every model was fitted using five-fold cross-validation (see for example Stone, 1974) to eliminate drastic overfitting, and the best model was chosen based on cross-validation accuracy. For every city, the data was split randomly in half to produce disjunct datasets for the two predictive tasks. It would also be possible to perform the estimation algorithm several times for each city with different data splits and use the median of the estimates to reduce the potential effect of a single "unrepresentative" split but I opted not to use this method due to its high computational costs .

A single city analysis – Madrid

To see how the outlined analysis looks in practice, I present it here for the case of Madrid. Madrid is in no way a special case in this analysis – it was chosen as it presents a good opportunity to contrast its results with the entire set of results, as we will see later. There are 10,990 Airbnb listings (after the previously described data preparation process) in the Madrid data for which I was able to identify the gender of the host. Most prices are under 200 euro a night with a few outliers, and there is a similar number of male and female hosts with a substantial number of couples as well.

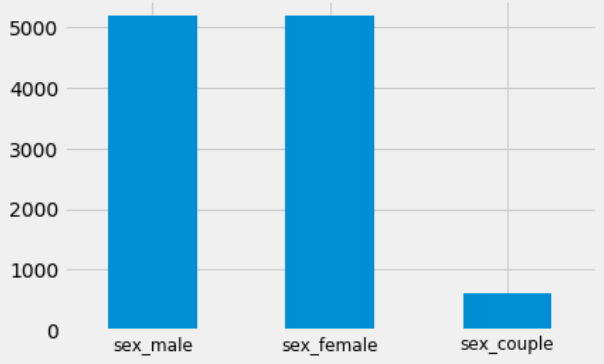


Figure 1: The number of Airbnb listings by host gender in Madrid

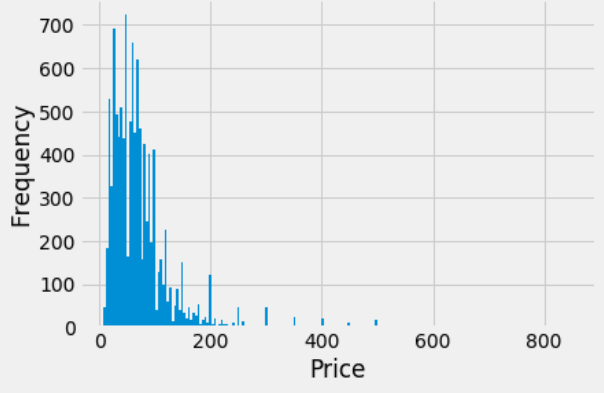


Figure 2: The distribution of Airbnb prices in Madrid

The distribution of prices looks similar for every gender, but couples seem to have slightly higher priced offerings. Professional hosts or hosts with more than one apartment on the site do not seem

to offer higher prices in general. (Prices higher than 300 euros are excluded from Figures 3 and 4 for easier interpretability.)

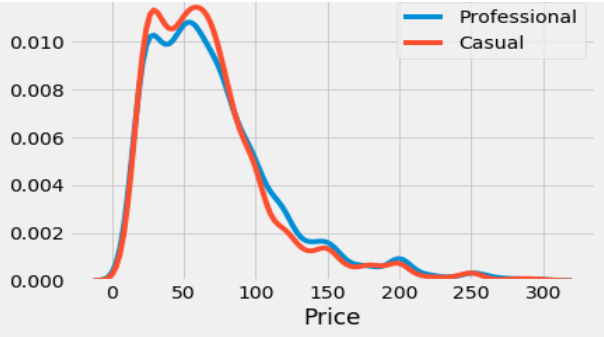


Figure 3: The distribution of Airbnb prices by host professionality in Madrid

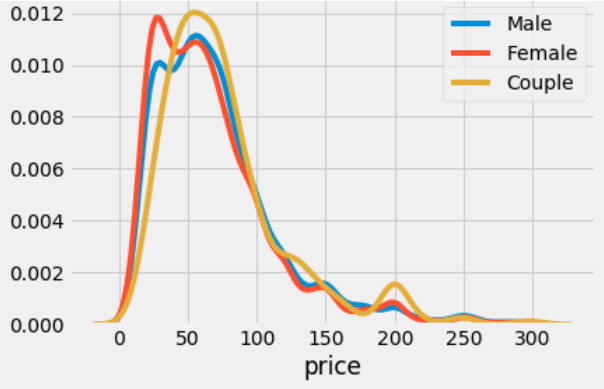


Figure 4: The distribution of Airbnb prices by host gender in Madrid

One can also observe clear spatial differences in mean prices. Offerings in more expensive neighborhoods could have more desirable attributes on average, but a purely spatial price effect is also plausible The patterns of the average prices on Figure 5 point out that these spatial differences are not only present but can take forms which linear functional forms on geographic coordinates are ill-suited to capture.

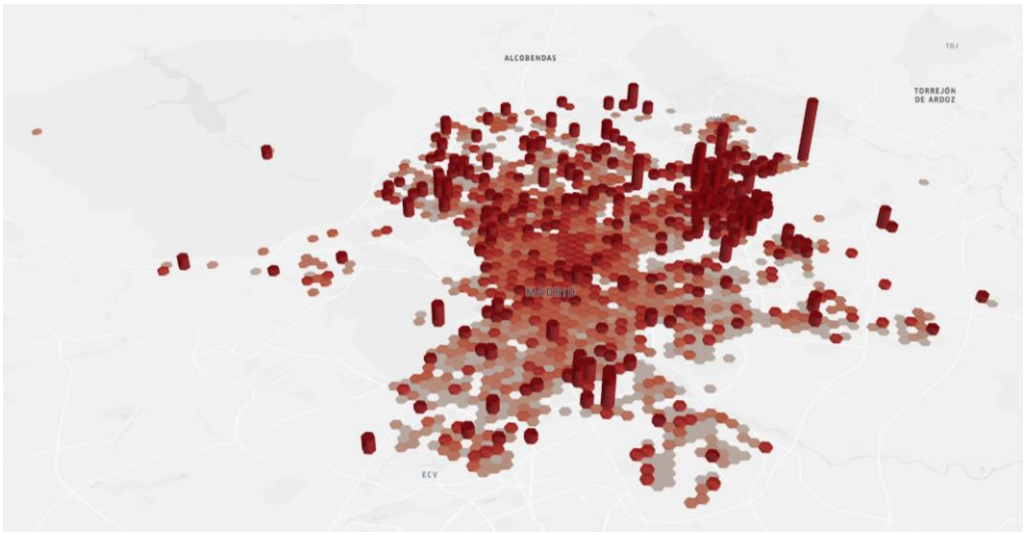


Figure 5: The spatial distribution of Airbnb prices in Madrid

– deeper red colors represent higher prices

Of course, both professional and casual hosts and hosts of different genders might have systematically different apartments offered on the site. It is necessary to control for these differences to understand how these attributes affect prices. My first approach to solve this problem is a simple linear regression with all apartment properties included. The dependent variable here is the log-price of offerings and the unit of observation is an individual property offering on the Airbnb site. This means there are 10,990 observations, each corresponding to a listed apartment at the time of data collection. Every other variable described in the data section is included as an independent variable[2], but no quadratic terms or interactions are included, except for the interactions between host gender, cohabitation and professionality. Table 3 summarizes the results of such a regression (amenity dummies and neighborhood dummies are excluded here as there are so many of them that they would make the table hard to navigate).

2 These variables are the following. As controls: number of guests accomodated, number of bathrooms, number of bedrooms, number of guests included in the price, number of reviews, averagre review score, number of reviews per month, a dummy indicating if the hosts identity is verified, latitude of the property, longitude of the property, number of features listed for the offering, a dummy for every amenity present in the dataset (after data preparation), a dummy for every neighbourhood present in the dataset (after data preparation). As the variables of interest: a dummy indicating a female host, a dummy indicating a couple host, a dummy indicating professionality, a dummy indicating cohabitation, and every possible interaction of these four variables.

Dep. Variable:	log_price	F-statistic:	77.57
Model:	OLS	Prob (F-statistic):	0.00
No. Observations:	10990	Log-Likelihood:	-5960.2
R-squared:	0.616	AIC:	1.237e+04
Adj. R-squared:	0.608	BIC:	1.400e+04

	Coefficient	Standard Error
accommodates	0.0967***	(0.0039)
bathrooms	0.0602***	(0.0085)
bedrooms	-0.0159***	(0.0026)
beds	0.0119***	(0.0038)
guests_included	0.0036	(0.0037)
number_of_reviews	-0.0005***	(0.0001)
review_scores_rating	0.0038***	(0.0005)
reviews_per_month	-0.0229***	(0.0030)
host_identity_verified	-0.0020	(0.0091)
latitude	1.5178***	(0.5787)
longitude	-0.4566	(0.5196)
Number of features listed	-0.0003	(0.0107)
professional	-0.0042	(0.0153)
live_together	-0.5412***	(0.0203)
pro_x_together	-0.0165	(0.0256)
pro_x_female	0.0030	(0.0213)
together_x_female	-0.0117	(0.0249)
pro_x_couple	-0.0437	(0.0457)
together_x_couple	0.0853	(0.0781)
pro_x_together_x_female	-0.0376	(0.0351)
pro_x_together_x_couple	-0.1184	(0.0971)
const	-59.3268**	(23.4848)

Omnibus:	3125.151	Durbin-Watson:	1.763
Prob(Omnibus):	0.000	Jarque-Bera (JB):	17139.945
Skew:	1.257	Prob (JB):	0.00
Kurtosis:	8.578	Cond. No.	1.13e+18

Table 3: Madrid OLS results

None of the gender-related variables have a significant effect (based on commonly used significance level thresholds) according to this regression. If this is a sufficient method to control for differences between apartments, then we should conclude that the gender of the host bears no

effect on the price of an Airbnb. A strict linear form, however, could lead to a heavy bias in this case as it is natural to suspect that prices might be related to apartment properties in ways far from linear. The use of the double machine learning method aims to solve this problem. Before discussing the results of DML, I briefly outline the Lasso results as well.

I first ran a Lasso regression – using the same specification as in the case of the regular linear regression discussed above – to find which variables end up with a coefficient different from zero. Then I ran an OLS with only those variables[3] to obtain valid standard errors. Table 4 summarizes the results from this second regression. None of the gender-related variables are included in this second stage, confirming the OLS results pointing to no significant gender-based effects.

Dep. Variable:	log_price	F-statistic:	969.8
Model:	OLS	Prob (F-statistic):	0.00
No. Observations:	10990	Log-Likelihood:	-8010.7
R-squared:	0.443	AIC:	1.604e+04
Adj. R-squared:	0.442	BIC:	1.611e+04
	Coefficient	Standard Error	
accommodates	0.2038***	(0.0025)	
bedrooms	-0.0049***	(0.0019)	
number_of_reviews	-0.0003***	(0.0001)	
review_scores_rating	0.0056***	(0.0006)	
reviews_per_month	-0.0190***	(0.0031)	
latitude	2.7957***	(0.2311)	
Shampoo	0.1425***	(0.0117)	
Free street parking	-0.3333***	(0.0145)	
const	-110.1157***	(9.3404)	
Omnibus:	1221.832	Durbin-Watson:	1.781
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4044.680
Skew:	0.562	Prob (JB):	0.00
Kurtosis:	5.751	Cond. No.	2.30e+05

Table 4: Madrid Lasso results

3 In this case, these remaining variables are the following: number of guests accomodated, number of bedrooms, number of reviews, averagre review score, number of reviews per month, latitude of the property, a dummy indicating the availability of shampoo at the offering, and a dummy indicating the availability of free parking at the offering.

To produce the DML estimates, I trained gradient boosting models to predict host gender and the natural logarithm of prices, according to the procedure discussed in the "Methods" section. The gender classifier model has a 0.83 average precision and recall, and a 0.41 R-squared statistic. The price prediction model has an R-squared statistic of 0.79998. In a sense, this gain in R-squared illustrates the whole point of using the DML method, as it shows how much better the relationship between price and other attributes can be captured by the model I chose to use than with a strictly linear form.

There is a caveat to this, however. With a universally flexible model and on a limited dataset, one could obviously always achieve a perfect fit without being able to generalize the results outside the dataset. To avoid this overfitting problem, I use the cross-validation method and regularization parameters described in the "Methods" section.

After training these models and obtaining their predictions, the prediction errors are used to estimate the gender effects. In the case of prices, the interpretation of the prediction error is straightforward. In the case of gender, a separate error is defined for both female hosts and couples. The error is one if the actual gender of the host is female (couple) while the predicted is not, minus one if the predicted gender of the host is female (couple) while the actual is not, and zero otherwise. (Technically, this error captures how much more or less a host is a couple (or female) than it would be expected based on their Airbnb offering.)

The distribution of both errors look regular (see Figures 6 and 7): the price error distribution exhibits a regular Gaussian shape, and most gender errors are zero (around 85% for female and

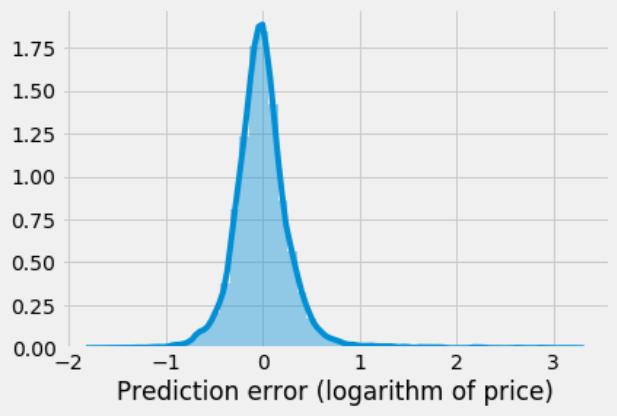


Figure 6: The distribution of price prediction errors

98% for couple hosts). As couples are less common among hosts, there are naturally fewer predicted hosts as well.

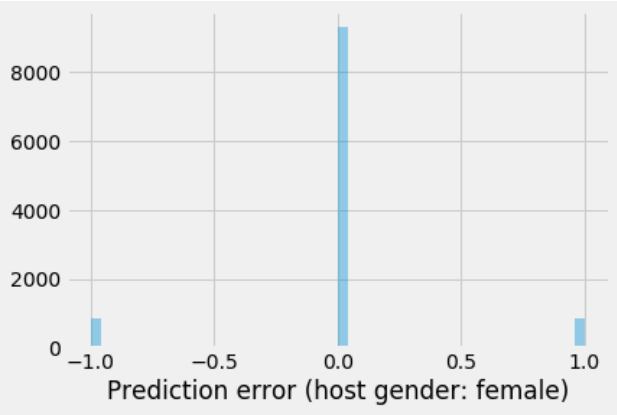


Figure 7: The distribution of gender prediction errors

In this final DML stage, gender effects ($\hat{\Theta}$) are estimated according to the formula:

$$\hat{\Theta} = arg\ min\ E[(\tilde{\mathrm{Y}} - \tilde{\mathrm{D}} * \Theta(X))^2]$$

where $\tilde{\mathrm{Y}}$ and $\tilde{\mathrm{D}}$ are the prediction errors. I specify the form of $\Theta(X)$ as:

$$\Theta(X) = \alpha_0 + \alpha_1 * professionality + \alpha_2 * cohabitation + \alpha_3 * (professional\ AND\ cohabiting)$$

(The last variable, "$professional\ AND\ cohabiting$" is an interaction.) These additive effects practically constitute a two-by-two matrix for professionality and cohabitation, two binary factors potentially determining gender effects. The constant is the estimate of the gender effect for a host who is not professional (only rents one property) and who does not live in the rented apartment herself (and therefore whose guests will not share it with her). Selectively adding the other three effects produces the estimate for any of the four possible combinations of relevant host attributes.

The final stage regression includes the prediction errors of log-prices as the dependent variable and the prediction errors of gender as the independent variable, with the coefficient $\Theta(X)$ specified as above, while the unit of observation is still the individual offering listed on the site:

$$\tilde{\mathrm{Y}} = \Theta(X) * \tilde{\mathrm{D}}$$

In the final stage linear regression for female hosts (see Table 5), only one of these four coefficients are significant at the 10% level (although the corresponding coefficient has a p-value of 0.01, so it is also significant at any level higher than 1%). The entire model, as measured by the F-statistic,

is significant at the 5% level, but not at the 1% level. This suggests that there might indeed be some relationship between prices and host gender even after controlling for other factors.

The interaction coefficient suggests that female hosts who are professional and live together with their guests set prices 5.6% lower than other hosts, after controlling for other factors. This is a very specific subset of hosts and there is no clear interpretation for why such an effect could exist. A possible explanation is that female hosts price less aggressively, but only when they price a cohabiting property "in comparison to" other properties of their own – so only when they also rent other properties (and are therefore professional by the definition used here). This is, of course, possible, but is a somewhat convoluted explanation. Other than this effect, I cannot confidently conclude – based on these DML results – that the host being female has an effect on the prices of Airbnb's in Madrid

Dep. Variable:	pred_error	F-statistic:	3.209
Model:	OLS	Prob (F- statistic):	0.0223
No. Observations:	1820	Log-Likelihood:	210.02
R-squared:	0.005	AIC:	-412.0
Adj. R-squared:	0.004	BIC:	-390.0

	Coefficient	Standard Error	t-value	p-value	95% Confidence Interval	
professional	0.0085	0.013	0.649	0.517	-0.017	0.034
live_together	0.0071	0.013	0.536	0.592	-0.019	0.033
pro_x_together	-0.0568	0.022	-2.580	0.010	-0.100	-0.014
const	0.0050	0.008	0.641	0.522	-0.010	0.020

Omnibus:	315.499	Durbin-Watson:	2.012
Prob (Omnibus):	0.000	Jarque-Bera (JB):	4670.901
Skew:	-0.332	Prob (JB):	0.00
Kurtosis:	10.820	Cond. No.	5.86

Table 5: Madrid DML results - female hosts

In the case of couples (see Table 6), three of the four estimated effects are significant at the 10% and two at the 5% level. What these results suggest is the following: couples charge 7 percent more, but only if they are professional and 12 percent more if they rent a property where they also live themselves. This is in line with my hypotheses that having a couple as the host might be worth

a premium on the Airbnb market, and that this premium is on the one hand higher when guests share the apartment with the hosts and on the other hand is more effectively priced in by professional hosts.

It is important to note that these are results after controlling for professionality and cohabitation – so, for example, shared apartments rented by couples might still be cheaper on average than entire, unshared properties, but are more expensive than shared apartments on offer by other hosts. Couples who are both professional and rent out a shared apartment only charge 4 percent more (the sum of the three coefficients disregarding the constant that is not significant), however. This is similar to the previous case, and once again might be explained by the hosts undervaluing their shared property in comparison to their own other properties.

Dep. Variable:	pred_error	F-statistic:		2.116		
Model:	OLS	Prob (F-statistic):		0.0982		
No. Observations:	302	Log-Likelihood:		-38.466		
R-squared:	0.021	AIC:		84.93		
Adj. R-squared:	0.011	BIC:		99.77		
	Coefficient	Standard Error	t-value	p-value	95% Confidence Interval	
professional	0.0732	0.037	2.001	0.046	0.001	0.145
live_together	0.1190	0.054	2.198	0.029	0.012	0.226
pro_x_together	-0.1506	0.079	-1.897	0.059	-0.307	0.006
const	-0.0229	0.028	-0.808	0.420	-0.079	0.033
Omnibus:	169.240	Durbin-Watson:		1.698		
Prob (Omnibus):	0.000	Jarque-Bera (JB):		1842.488		
Skew:	2.056	Prob (JB):		0.00		
Kurtosis:	14.380	Cond. No.		6.98		

Table 6: Madrid DML results - couple hosts

The results of this single city analysis are not entirely conclusive. They suggest that some of the hypothesized effects do exist, but only in the case of couples – namely the positive one associated with having a couple as a host. This effect is higher for cohabiting properties and professional hosts, as it was hypothesized. There is another, unexpected effect for both couples and females

host (the negative one associated with the interaction variable) with no simple and convincing explanation. Both hypothesized effects for females seem to be nonexistent.

These conclusions are not absolutely clear and unequivocal – the significance levels of estimated coefficients do not reach 1%, which is a regularly used, albeit arbitrary threshold to assess how solid an empirical result is. Altogether, however, this single city analysis appears to confirm a subset of the initial hypotheses.

Repeating the same analysis for 60 other locations can shed a different light on these individual results as well. A researcher can pose the questions discussed here, produce estimates for a number of cities and cherry-pick one with the most significant and interesting conclusions. Results produced this way are not valid as there is a clear need for correcting significance thresholds to account for the multiple implicit significance tests executed in this case. The choice of an ideal correction method however might depend on multiple factors, including how the different "parallel" regressions and significance tests are related.

One reason to execute this same analysis over multiple cities is to obtain a sample of estimated coefficients instead of single values. One, somewhat extreme, case of the relationships of the "parallel" regressions is to consider the data from all 61 cities as produced by the same single underlying data generation process. It is implausible in this pure form but is nonetheless a potentially insightful perspective. The 61 results can be interpreted as 61 independent "experiments" about entirely independent processes and effects, or, in this latter case, as 61 "experiments" about the same process.

In reality, the truth is almost certainly between these two extremes. (Although truth is arguably a dubious abstraction in this statistical context.) The question is the same for every location and in many aspects the sample of locations is fairly homogeneous, as it is drawn from the most popular tourist destinations of more wealthy countries. Therefore, it would be naive to claim that the theoretical data generating processes behind the data are completely independent and one does not contain information about the others. Assuming an identical data generating process is, as I stated, probably also a simplification, but I will analyze the results from this perspective as well to gain a better understanding of the results.

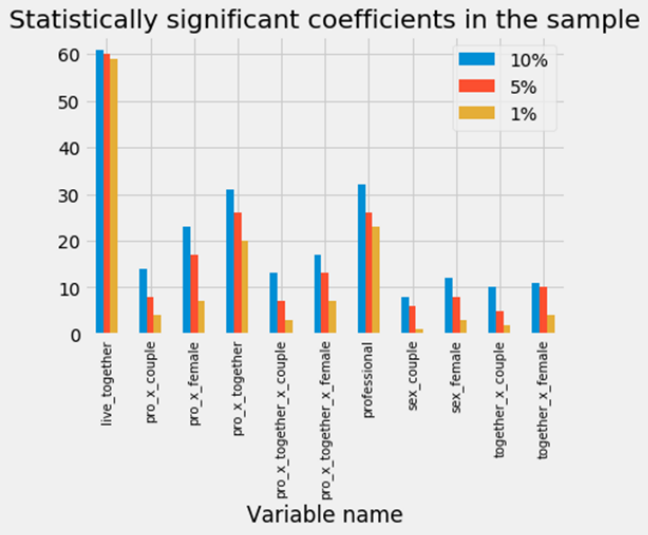
Results over the entire sample of cities

I analyze the results from the 61 city analyses from two parallel perspectives. I look at them one-by-one to assess what kind of effects are present in each city if I consider them independently, and which effects are more common than others. I also look at the results as 61 realizations of the same underlying process (as articulated in the previous section) and draw subsequent conclusions about said process. I also try to assess how plausible it is that the results are produced by the same process. To do this, I look at the sets of 61 coefficients individually and examine whether they could be from a common distribution with specific properties. This is a limitation in the sense that I only study if single coefficients are constant across cities, not the complete data generating process in its entirety.

Linear regression results

First, I look at how many of the coefficients in question are significant at the established significance levels of 10%, 5% and 1%. (This is summarized by figure 8.) Cohabitation – as it is to be expected – is virtually always highly significant and has a negative impact on price in all 61 locations. The most commonly significant coefficients behind it are professionality and, somewhat surprisingly, the interaction between professionality and cohabitation. I expected professionality to increase prices, and it is indeed the case in most cities according to linear estimates, but not all (38 out of 61 overall but 17 out of 23 the coefficients significant at 1%). The interaction between professionality and cohabitation is significant at the 1% level in 20 cases and has a negative sign in 17 of those. This effect came up in the Madrid analysis as well, where I outlined a potential, albeit debatable explanation.

Looking only at gender-related variables, they are significant substantially less often. It is also important to note that when looking at several coefficient's significance, one would expect some of them to appear significant simply by chance even if all of them were zero in reality. If I apply the Bonferroni correction, or the slightly less conservative Šidák correction to the results, I observe that gender-specific variables almost never reach statistical significance (see Figure 9).



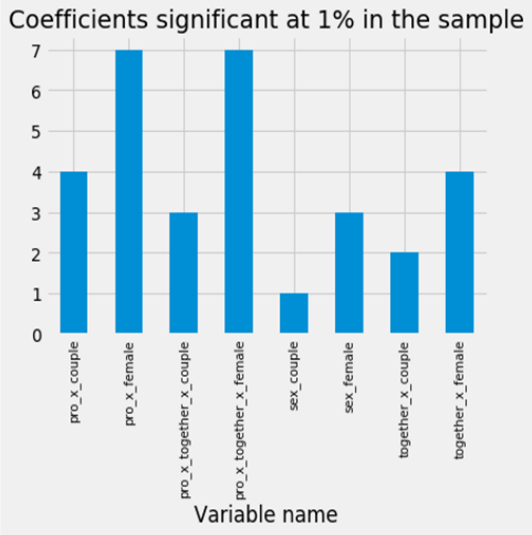
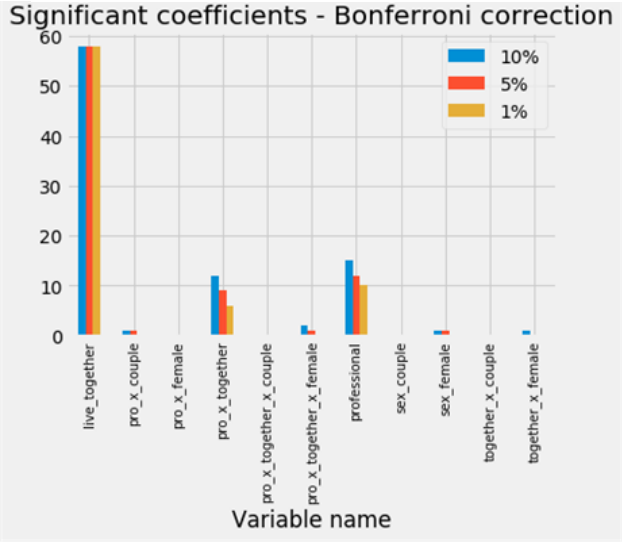


Figure 8: The number of statistically significant OLS coefficients



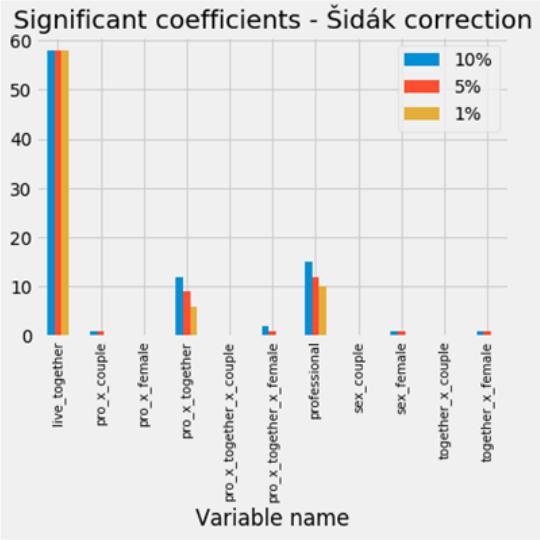


Figure 9: The number of significant OLS coefficients after multiple comparison corrections

Based on these analyses I conclude that, according to linear regression results, we cannot evince any gender-specific effects on Airbnb prices in general. One might consider the few significant results as special cases where there are indeed such effects, or mere statistical outliers. (These "exceptions" significant at 5% even after corrections are: Paris where professional couples charge 7% less, New Jersey where professional and cohabiting female hosts charge 30% (!) more, New York where female hosts charge 2% less and Dublin where cohabiting female hosts charge 2% less.)

Looking at significance levels without knowing the sign and magnitude of coefficients of course provides limited understanding, so now I turn to studying the distribution of coefficients. The histograms on Figure 10 illustrate those distributions: the blue curves are fitted density functions and the black ones are fitted normal densities. The green lines are at zero and the red lines are at the mean of the coefficients.

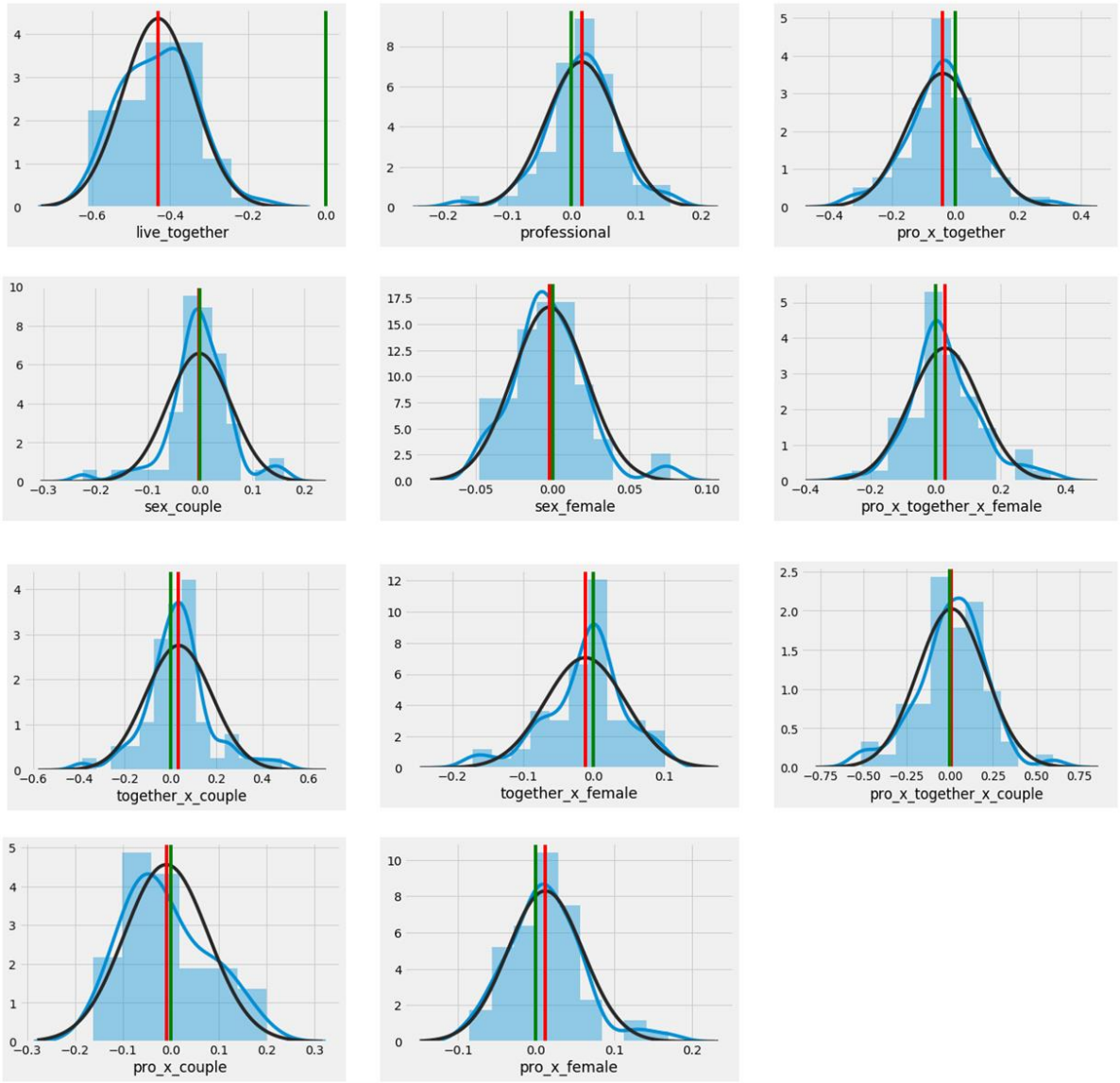


Figure 10: The distributions of OLS coefficients

(blue line: estimated kernel density; black line: fitted normal density; green line: zero; red line: sample mean)

Four of the coefficients have a mean that is significantly different from zero at the 5% level, and two other means are different from zero at the 10% level. (As determined by using a t-test.) The

former four are cohabitation (the mean is -0.43 with a p-value of practically zero), professionality (the mean is 0.015 with a p-value of 0.035), the cohabitation-professionality interaction (the mean is -0.039 with a p-value of 0.009), and the cohabitation-professionality-female interaction (the mean is 0.029 with a p-value of 0.042). The latter two are the professional-female interaction with a coefficient of 0.012 and the cohabitation-couple interaction with a coefficient of 0.035.

To assess if the coefficients could be from a single distribution, I tested their distribution for normality and symmetry. Any empirical distribution of coefficients is clearly compatible with them coming from a single arbitrary distribution, but not with them coming from a normal, or less restrictively a symmetric (and unimodal) distribution. I test for symmetry around both the empirical mean coefficient and zero, using the Wilcoxon signed-rank test and test for normality using the Shapiro-Wilk test. I test if the distributions are normal or at least symmetric and unimodal to assess if they could be from a single sampling distribution of an OLS coefficient in which case the true mean of this sampling distribution is the true parameter value.

At the 5% level, the null hypothesis of normality is rejected for four coefficients: “female”, “couple”, the cohabitation-female interaction and the cohabitation-couple interaction. Using the same p-value threshold, none of the distributions are symmetric around zero but all are symmetric around their mean. Overall, the analysis does not rule out that the coefficients come from the same distribution with a common mean, which is also the true parameter value in this case. This true parameter value is then clearly significant for cohabitation and once again for the cohabitation-professionality interaction – the significance of other coefficients is less clear based on the p-values reported above (as I stated, some more are significantly different from zero at the 5% level, for example).

Funnel plots are another tool to examine whether significant coefficients (as determined by the traditional p-value thresholds) are the results of chance or true effect. The share of significant results and their symmetry or asymmetry around zero can help answer the aforementioned questions. These plots on Figure 11 confirm the previous conclusions and strengthen the case for the cohabitation-professionality-female interaction to be considered as significant. In the case of this triple interaction every coefficient significant at 1% is greater than zero and more significant coefficients are generally concentrated on the positive side.

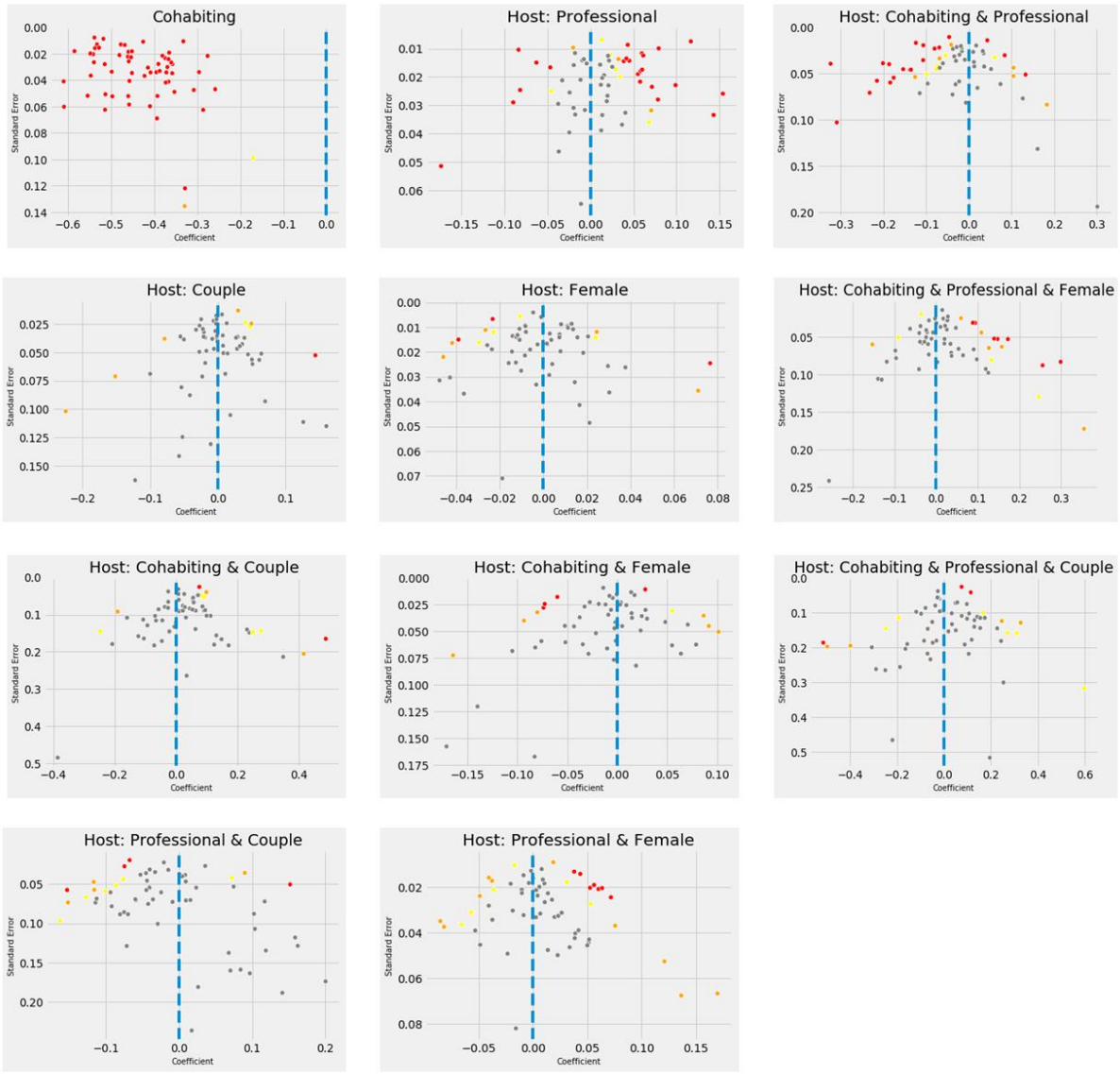


Figure 11: Funnel plots of OLS coefficients

(red: significant at 1%, orange: significant at 5%, yellow: significant at 10%, grey: all others)

Lasso results

Lasso sets a higher thresholdfor a coefficient to remain nonzero but consequently a remaining coefficient is more likely to be highly significant in the second regression (with a limited number of independent variables). This second effect is increased by the lower number of controls in the second regression, which controls could be correlated with the remaining variables. This is reflected in the Lasso results, as summarized by the number of significant coefficients (see Figure

12). Most variables of interest are rarely or never significant, but the significant ones mostly remain significant even after the Bonferroni correction (see Figure 13).

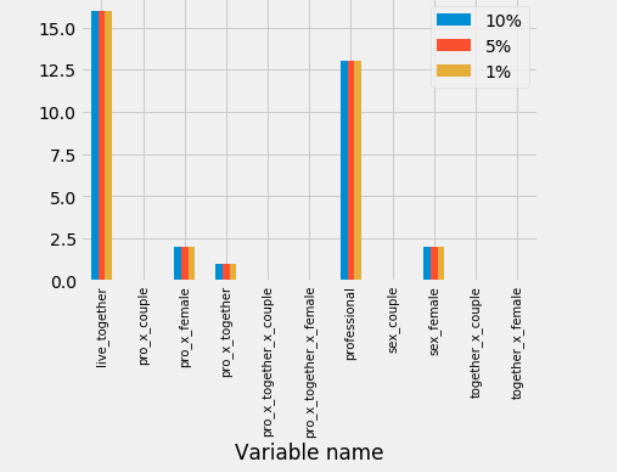


Figure 12: The number of statistically significant Lasso coefficients

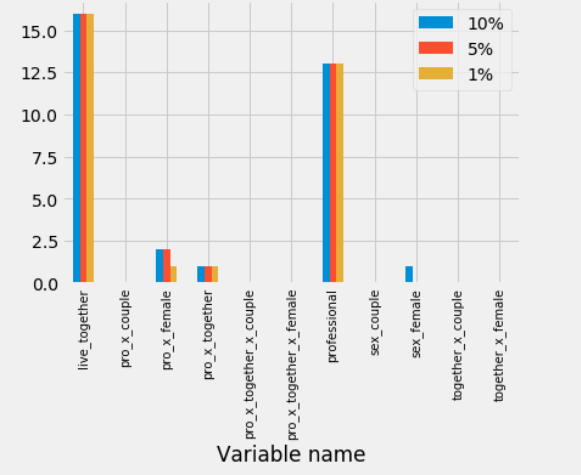


Figure 13: The number of significant Lasso coefficients after multiple comparison corrections

Cohabitation and professionality are significant in many cases, while the other variables are generally insignificant. The exceptions for the effect of a female host are Menorca and Jersey City (with effects of positive 9% and 5%, respectively). For the professionality-female interaction they are Tokyo and Ghent (16% and 13%) and for the cohabitation-professionality interaction it is Hong Kong (-34%). Taken together, these results do not contradict the previous conclusions that gender has no effect on Airbnb prices.

Of course, Lasso does not address the potential problems arising from nonlinearities – for this, I use the double machine learning method. This can help determine whether the linear regression results were biased and whether the conclusions based on linear models are correct. I outline these results in the following subsection.

Double machine learning results

With the DML method, I do not estimate the effects of cohabitation and professionality, as they are controlled for during the machine learning modeling phase, which does not result in any causal estimates. The other eight variables discussed so far, on the other hand, have their analogous counterpart in this analysis.

Looking at the number of significant coefficients (by traditional standards), there are less of those here than in the linear case (see Figure 14). If I apply the Bonferroni correction, only one coefficient remains significant (and only at the 5% level); this estimate suggests that couples charge 10% more for their properties in London (but other estimates suggest this effect is only present for unprofessional couples). Among the individual results, this is the only case where one can confidently argue for the existence of a true effect.

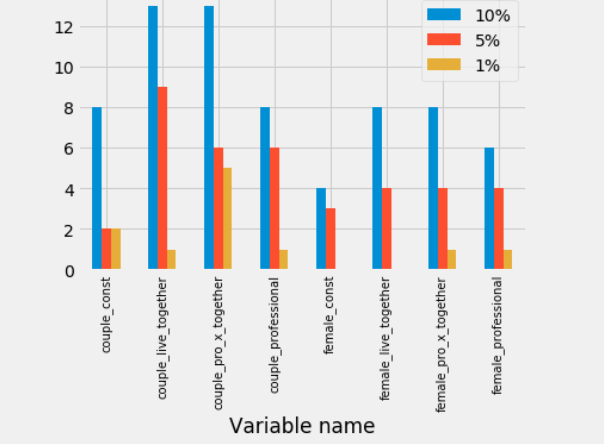


Figure 14: The number of statistically significant DML coefficients

Other effects significant at the 1% level without corrections are a 4% effect for couples in Munich, a 0.9% effect for cohabiting couples in Ghent, effects of 8%, 24%, -39%, -34%, 35% for cohabiting professional couples in Austin, Milan, Berlin , Belize and Naples, respectively; a -9% effect for professional couples in London (basically neutralizing the positive effect for couples in general),

a -6% effect for cohabiting professional female hosts in Athens and a -9% effect for professional female hosts in Washington.

Now I turn to examining the distributions of the estimated coefficients. The means of these distributions are not statistically different from zero in any of the eight cases (as determined by a t-test). According to the Shapiro-Wilk and Wilcoxon signed-rank tests, four of these distributions are not statistically different from normal and are symmetric, while for four others, these hypotheses are rejected at the 5% level. These latter four are the coefficients for professional couples, female hosts, cohabiting female hosts and professional female hosts. These therefore are unlikely to all come from the sampling distribution of an OLS coefficient and so unlikely to correspond to one single underlying true parameter value.

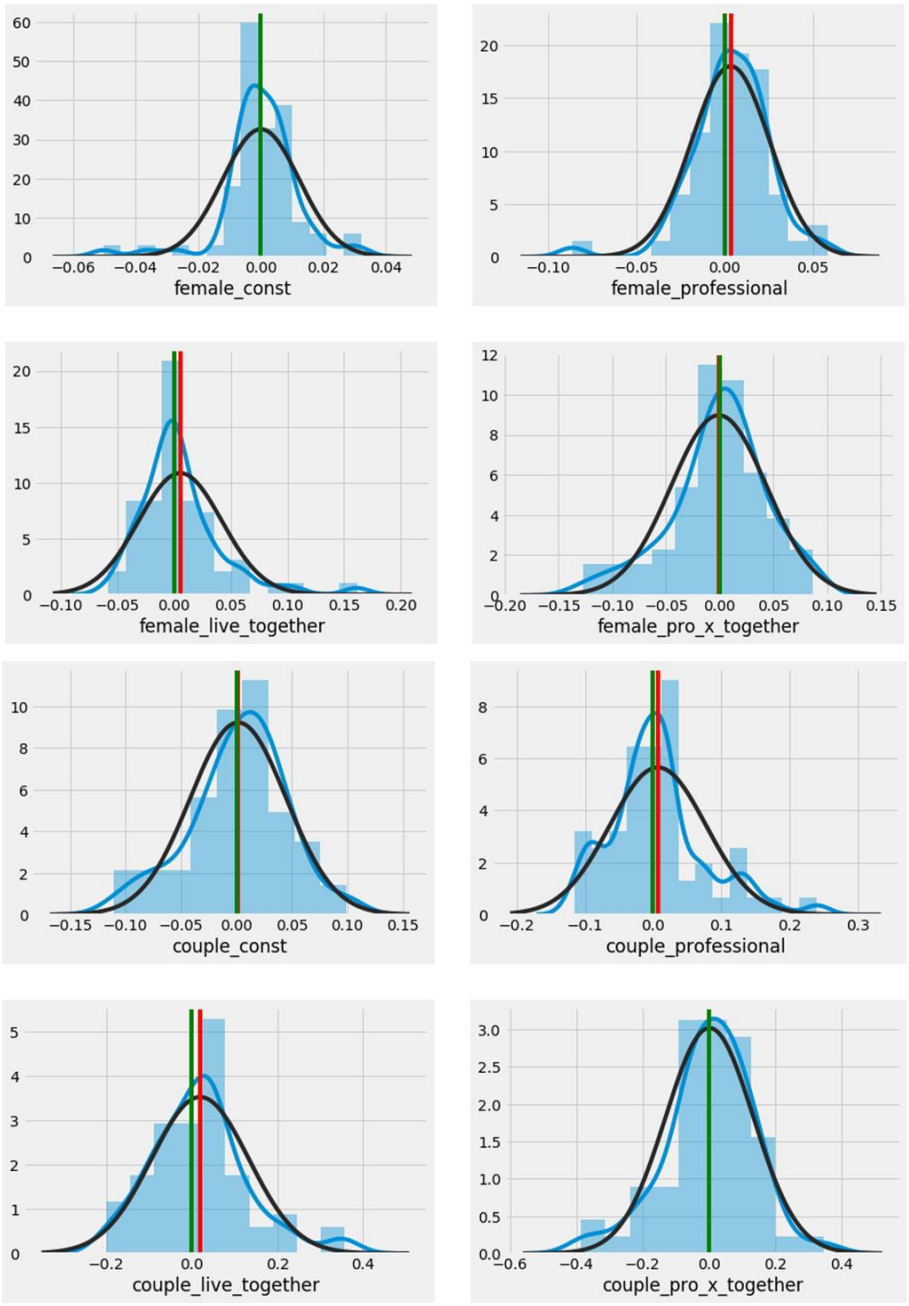


Figure 15: The distributions of DML coefficients

(blue line: estimated kernel density; black line: fitted normal density; green line: zero; red line: sample mean)

The results suggest that the estimates in these four are unlikely to come from a single distribution. Based on the distributions as visualized on figure 15, these "outlier" coefficients are probably negative for female hosts and positive for the three others. This is in line with my original hypotheses, but these results are not consistent across cities and with very few exceptions are not unambiguously significant if taken individually. The funnel plots (figure 16) also do not point to

the presence of any clear effects as the distribution of estimated coefficients is generally more symmetric than in the case of linear regressions.

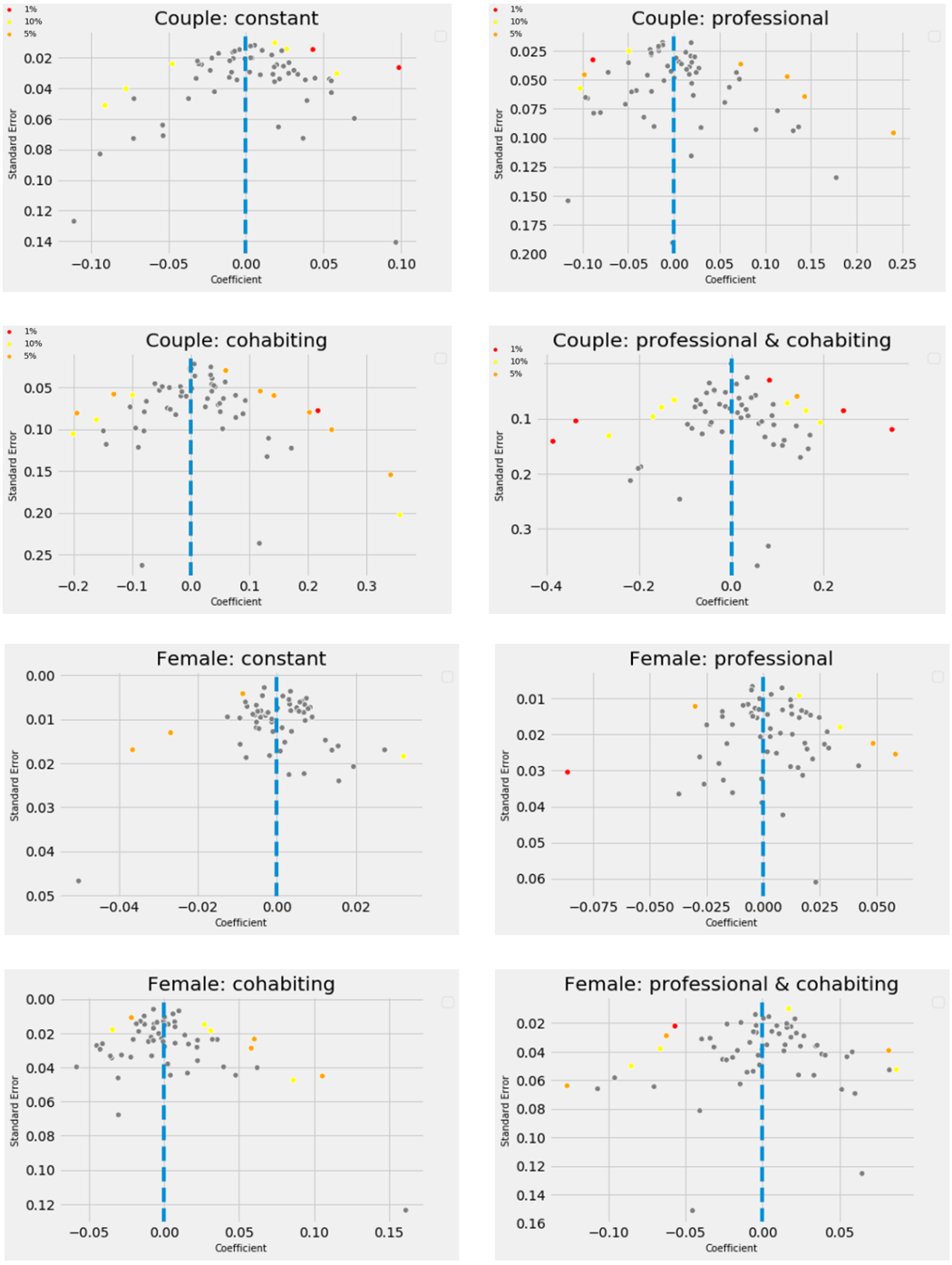


Figure 16: Funnel plots of DML coefficients

(red: significant at 1%, orange: significant at 5%, yellow: significant at 10%, grey: all others)

To most hosts, of course, it is not single coefficients but sums of coefficients that apply: the overall estimated effect for professional couples, for example, is the sum of the "couple" and the "couple and professional" coefficients. I do not test for the significance of these sums or "natural coefficients" but look at their distributions to better understand the estimated effects (see Figure 17). (Two of these are of course the same as on Figure 15.)

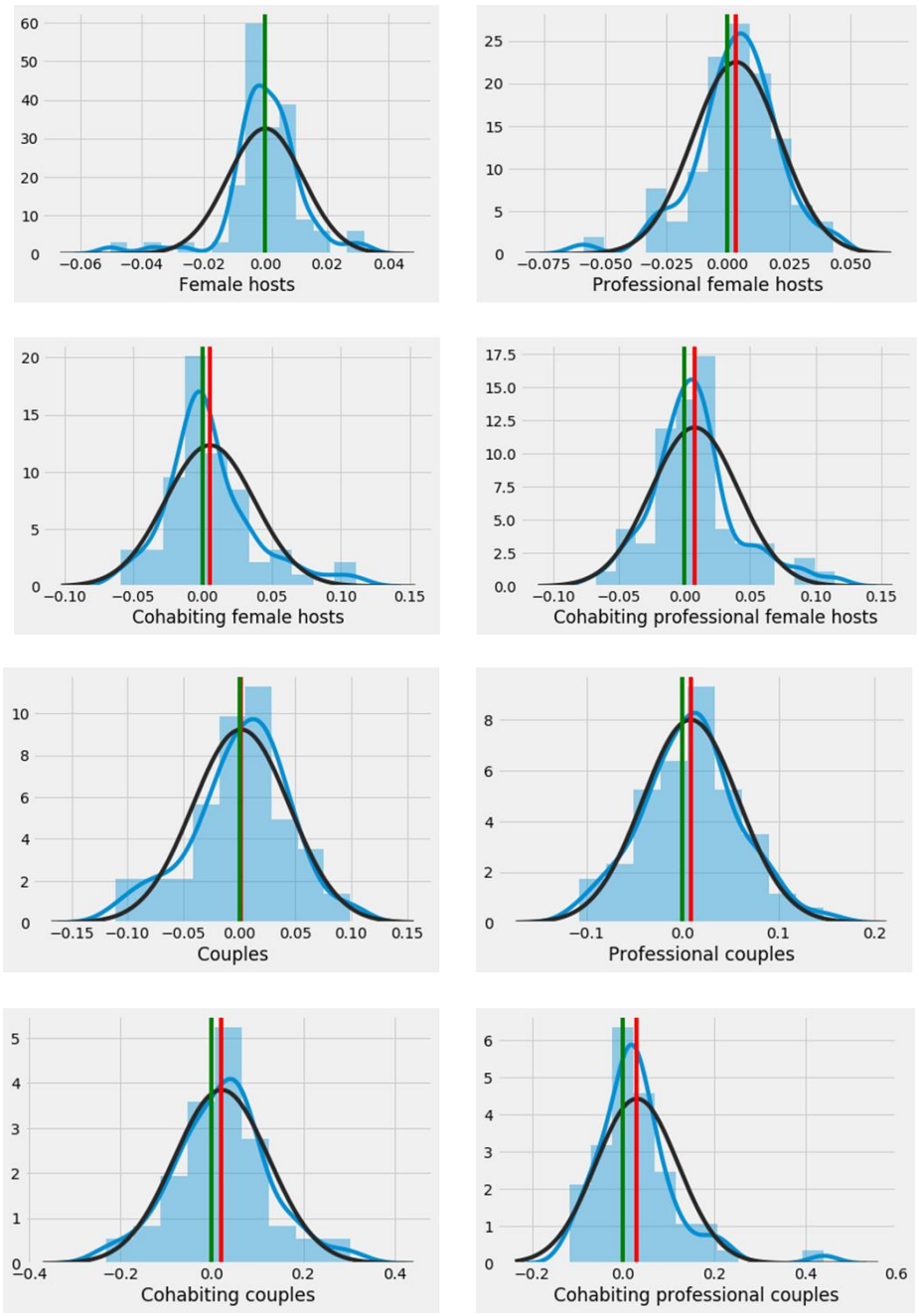


Figure 17: The distributions of DML effects

(blue line: estimated kernel density; black line: fitted normal density; green line: zero; red line: sample mean)

Among these eight, the mean effect for professional cohabiting couples is different from zero at the 5% level (one other, the coefficient for cohabiting professional female hosts is significant at the 10% level). If I assume that coefficients are the same for all locations and the estimated coefficients come from a common distribution, then this is the single one coefficient for which it can be said that probably there is a true underlying effect. The mean professional cohabiting couples is 0.03, corresponding to a 3% effect. (Similarly, the corresponding effect would be 0.8% for cohabiting professional female hosts.)

Generally, the DML method leads to less significant coefficients than the linear regression – to what extent this is due to better controlling for property attributes or using less information present in the data is not analyzed here. As the potential for nonlinearities is great in this context, I consider the DML results sounder and more reliable. Looking at these results one-by-one, the single effect we can be confident about is the 9% positive one for couples in London. (Which itself comes with a negative effect of similar size for professional couples which is however not significant after the Bonferroni correction.) This confidence of course depends on the choice of significance level thresholds and the correction method for multiple comparisons: I base my assessment on the 5% threshold and the Bonferroni correction.

Looking at the entire sample of coefficients, none of the mean coefficients are statistically significant from zero. In four cases the distributions suggest that all estimates could be from a single underlying distribution, in four others they point to the presence of "outliers". These seem to be negative for female hosts and positive for professional couples, cohabiting female hosts and professional female hosts.

The estimated effects are (at least in most cases) actually sums of estimated coefficients. Studying these sums, potential positive "outliers" are observed for cohabiting professional couples, cohabiting female hosts and cohabiting professional female hosts, and negative ones for female hosts. In one case, the mean effect is statistically significant from zero (with a p-value of 0.012): a mean effect of positive 3% is observed for professional cohabiting couples. (I also perform multiple comparisons, of course, when testing the sample mean of coefficients against zero and therefore a correction could also be applied here – the Bonferroni correction would turn the 10% threshold into a 1.25% threshold and the 5% threshold into a 0.625% threshold. The mean coefficient for professional cohabiting couples would clear the former but not the latter.)

The analysis presents no solid evidence for most of the initial hypotheses. The only single effect that is strongly significant is the 10% positive effect for couples in London. The coefficient distribution asymmetries suggest that there are cases where a female host has a negative effect on price, which is moderated or even reversed by the host being professional or cohabiting. There is clearer evidence that professional couples can charge a premium, but only in a cohabiting context (of course, other explanations are also possible for this effect). This result arises from considering the estimates as a sample from an underlying distribution, and therefore the effect cannot be pinned down to particular locations. These results are in line with the initial hypotheses but are both limited and in most cases ambiguous.

The comparison of OLS and DML estimates

Both the OLS and the DML results were discussed previously in this paper and some of their qualities were contrasted in order to answer the research questions. In this section I compare these two sets of results directly to better understand the differences between these two estimation methods. To illustrate these differences the point estimates from the two methods are plotted on Figure 18 with fitted regression lines and the corresponding 95% confidence intervals.

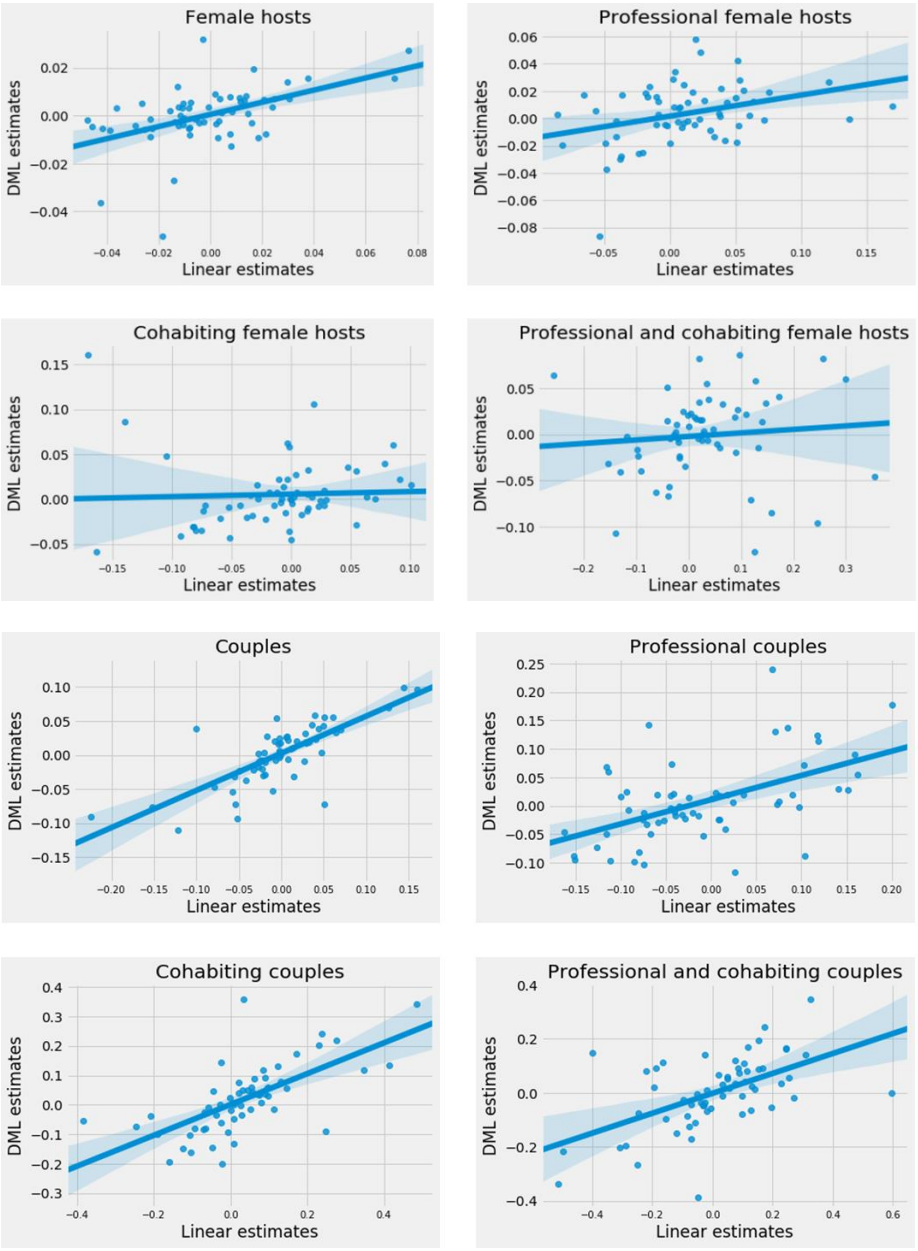


Figure 18: Scatterplots of DML and OLS coefficient estimates

Generally, the DML estimates exhibit a lower variance than OLS estimates. This is potentially attributable to better precision achieved via controlling for the non-gender-related determinants of Airbnb prices more precisely. An alternative explanation would be that DML utilizes less information present in the data and is therefore more likely to result in estimates closer to zero – but this does not follow from the theoretical properties of DML (Chernozhukov et al., 2016) and would point to some additional imperfections of the estimation strategy.

For this reason, I consider the first explanation more likely; in this case, what we see is an evidence of DML's ability to produce more precise estimates. As the majority of estimated coefficients is not statistically different from zero – regardless of method – I interpret the DML coefficients being closer to zero as simply being closer to the true parameter and thus being more precise. (The mean of the estimated DML coefficients is always smaller than or equal to the mean OLS coefficient in absolute value.)

There are positive correlations between the estimates from the two methods, which is to be expected. It is more noteworthy that this relationship is much clearer and stronger for variables related to couples (especially in the case of cohabiting and triple interactions). This also points to the conclusion that gender-specific effects are not invincible for females but are to some extent present for couples. If the true effect for female hosts is indeed zero, then parameter estimates for this coefficient could just be random noise around zero and thus the two estimates could be uncorrelated. A positive correlation, however, is not only possible if there is a true effect which is to some extent captured by both methods but also if the two methods exhibit similar biases and are "wrong in the same direction".

The scatterplots of standard errors (Figure 19) also point to the DML results being more precise. The positive correlations are much more pronounced in this case but the DML standard errors are with very few exceptions lower than their OLS counterparts.

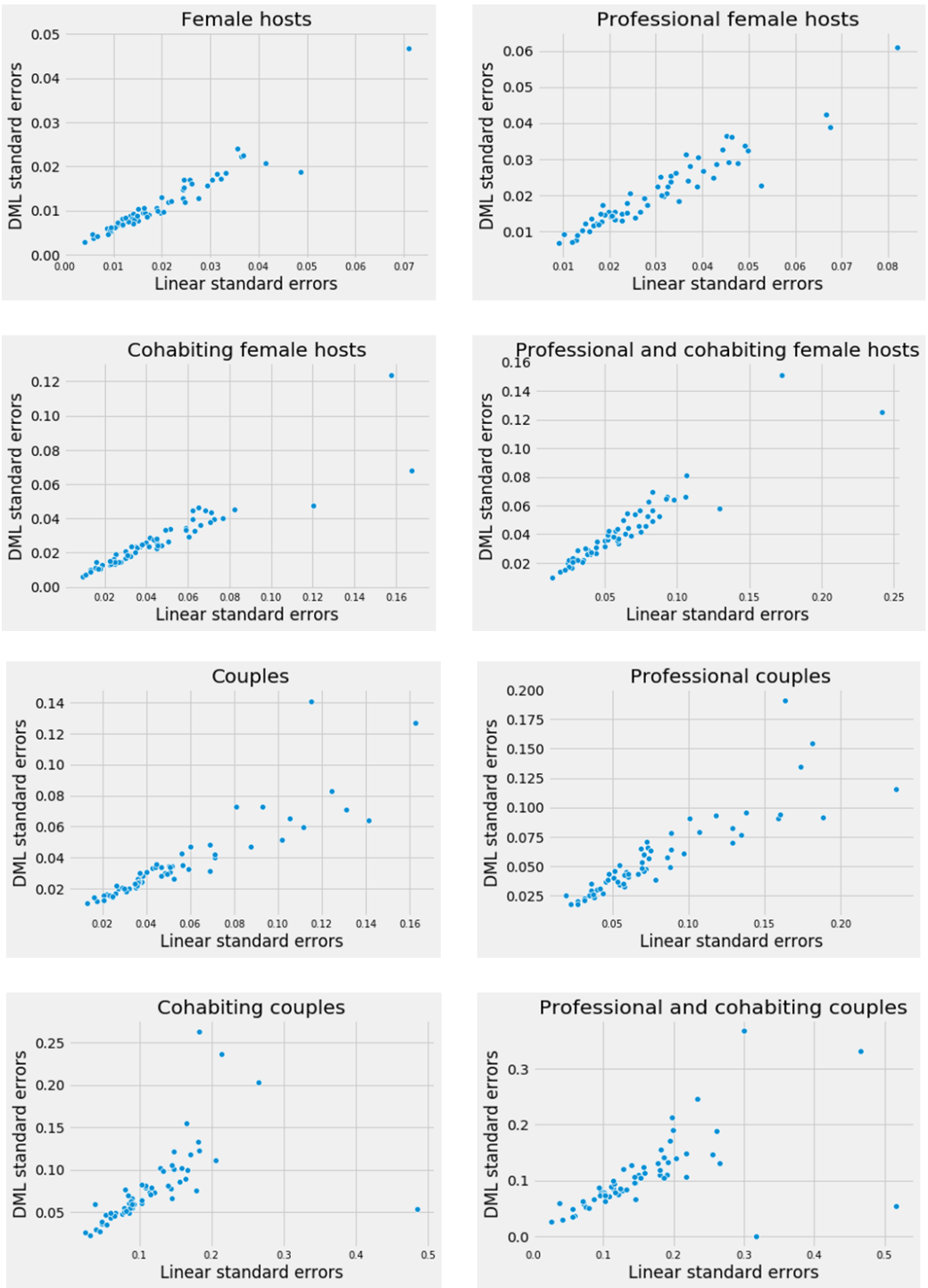


Figure 19: Scatterplots of DML and OLS standard errors

Conclusions

In this paper I examined how the gender of the host affects the prices of Airbnb offerings in 61 different locations. I used the "double/debiased machine learning" method (Chernozhukov et al., 2016) to answer this question as there is a high number of potential confounders present in this context and it would not be prudent to assume that these affect prices in a linear way – or in any other specific form that would be recognizable ex ante. I also use classical linear regression (and Lasso) to answer the same questions and compare the results of the two methods. I do so in order to present a case study to better understand their differences in a practical setting.

This contributes to the research on the relationship between Airbnb prices and personal host attributes in two novel ways. My study focuses on gender differences, which topic is much less developed than the effect of race in similar settings. Additionally, I studied how gender-specific effects vary with professionality and cohabitation – these factors would be relevant in the case of racial effects as well. The paper also accounts for the potential complex nonlinearities between prices and determining factors, an issue mostly neglected in previous research.

Estimating gender effects on Airbnb prices is complicated because there are different potential effects with opposite signs. My first hypothesis was that a female host has a negative effect on pries via gendered differences in pricing behavior – female hosts setting lower prices for identical properties on average. My second hypothesis was that a female host has a separate positive effect on Airbnb prices because guests prefer female hosts over male ones for safety reasons at least in some cases. Other similar channels could also be present, for example if guests see hosts of a specific gender as more trustworthy.

To disentangle these effects, I utilized three approaches. I included hosting couples in the analysis in addition to males and females. I hypothesized that the aforementioned positive effect is analogously present for couples but the negative one is not. This is a strict assumption that would allow me to precisely separate the two effects. Even without making this strict an assumption, including couples helps to understand the effects in question better.

I also studied the interaction of cohabitation (the host and the guest living together) and host gender. It is rational to assume that safety-related positive effects for females are higher in this case – if I make the stronger assumption that they are only present in this case, then it is once again possible to precisely separate the two effects. Finally, I included the interaction of professionality

(as proxied by the overall number of listings offered by a host) and gender. Here I made the assumption that professional hosts exhibit no gender differences in pricing, but the other effects are present in this case: this once again means that the two effects are precisely separateable. These methods allow for checking the effect sizes based on different assumptions and to see how the estimated effects change based on the assumptions – and to assess the consistency of the assumptions.

The OLS and DML methods produced similar results, but the results of the latter exhibit smaller variance. The coefficients are generally distributed around zero and for most combinations of the examined host attributes there are no consistent positive or negative effects.

Several individual coefficients were significant at the traditionally used significance levels but as a high number of parameters were tested against being zero simultaneously it is necessary to correct for multiple comparisons. After applying the Bonferroni correction, only one DML estimate was significant at the 5% level: this shows that the host being a couple has a 10% positive effect on the price of an Airbnb. Whether using the conservative Bonferroni correction is ideal is debatable (and so is the choice of p-value thresholds); therefore, there is no indisputable conclusion about the individual coefficients. Generally, however, I found no consistent effects this way.

I also analyzed the estimated coefficients together, as samples of observations – for all eight relevant coefficients. The most extreme interpretation of this perspective is to consider all 61 estimates as coming from a single underlying distribution with an expected value that is equal to a single underlying true parameter value. I examined the means of the distributions of the estimated coefficients and tested these distributions for normality and symmetry.

The mean of these samples was not significantly different from zero in any of the cases. For four of the distributions the assumptions of normality and symmetry were rejected (while they were not rejected in the other half of the cases). These four cases were the coefficients for professional couples, female hosts, cohabiting female hosts and professional female hosts. This means that these four samples are unlikely to come from a single common unimodal and symmetric distribution. Based on observing their empirical distribution, potential outliers that cause this asymmetry are likely to be negative for female hosts and positive for the three other categories. This result is in line with the initial hypotheses, but it is important to remember that these outliers are not individually strongly significant.

As I estimated heterogeneous effects with additive coefficients, the actual effects for many hosts would be the sum of some estimated effects. I analyzed the distributions of these more naturally interpretable sums as well, but in this case, I did not calculate standard errors for these summed coefficients. One of the summed or "natural" coefficients had a mean (0.03) statistically different from zero (with a p-value of 0.012). I interpret this as an average effect of positive 3% for professional cohabiting couples. A possible explanation for this effect would be that the host being a couple has a positive effect on price because it is preferred by guests if they share the property with the host(s), but this is only effectively priced in by professional hosts. The absence of a similar effect for female hosts could be caused by a balancing negative effect on the pricing side or by the guest preference being exclusive for couples.

The results of the analysis did not generally confirm the initial hypotheses. Only one very specific effect is observed consistently with the DML method, while the occasional presence of some other effects is likely but more ambiguous. Estimated effects are much more pronounced for couples then for female hosts, but even these effects are debatable. These tendencies are compatible with the initial hypotheses – larger couple effects could be the results of additional negative effects for females that bring the overall effect to somewhere around zero for them, for example. Based on the results, there is no evidence of the original hypotheses about female hosts.

The results suggest that some or all of the hypothesized effects are present for couples at least in a subset of locations, but further research would be necessary to precisely pin down or reject such effects. This could be achieved using more creative estimation strategies or focusing the research more on couples instead of female-male differences. Observing effects for only couples also have different implications in general. If being (part of) a couple is advantageous in this setting, then the reasons behind this phenomenon should be unambiguously identified and their potential presence in other spheres of life should also be examined.

References

- Airbnb, 2020. “How is the price determined for my reservation?”, website: https://www.Airbnb.com/help/article/125/how-is-the-price-determined-for-my-reservation, downloaded on: 2020.05.09.

- Buser, Thomas & Niederle, Muriel & Oosterbeek, Hessel, 2014. "Gender, Competitiveness, and Career Choices," The Quarterly Journal of Economics, Oxford University Press, vol. 129(3), pages 1409-1447.

- Buser, Thomas & Peter, Noemi & Wolter, Stefan C., 2017. "Gender, Competitiveness, and Study Choices in High School: Evidence from Switzerland," American Economic Review, American Economic Association, vol. 107(5), pages 125-130, May

- Buser, Thomas & Yuan, Huaiping, 2019. "Do Women Give Up Competing More Easily? Evidence from the Lab and the Dutch Math Olympiad," American Economic Journal: Applied Economics, American Economic Association, vol. 11(3), pages 225-252, July

- Chattopadhyay, Manojit & Mitra, Subrata. 2019. “Do Airbnb host listing attributes influence room pricing homogenously?,” International Journal of Hospitality Management. 81. 54-64. 10.1016/j.ijhm.2019.03.008.

- Chen, Tianqui, & Guestrin, Carlos, 2016. “XGBoost: A Scalable Tree Boosting System”, arXiv e-prints, arXiv:1603.02754.

- Chernozhukov, Victor & Chetverikov, Denis & Demirer, Mert & Duflo, Esther & Hansen, Christian & Newey, Whitney & Robins, James, 2016. "Double/Debiased Machine Learning for Treatment and Causal Parameters," Papers 1608.00060, arXiv.org, revised Dec 2017.

- Choudhary, Paridhi & Jain, Aniket & Baijal, Rahul, 2018. "Unravelling Airbnb Predicting Price for New Listing," Papers 1805.12101, arXiv.org

- Daisuke, Miyakawa 2019. "Shocks to Supply Chain Networks and Firm Dynamics: An Application of Double Machine Learning," Discussion papers 19100, Research Institute of Economy, Trade and Industry (RIETI)
- Edelman, Benjamin & Luca, Micahel, 2014. "Digital Discrimination: The Case of Airbnb.com," Harvard Business School Working Papers 14-054, Harvard Business School
- Ert, Eyal & Fleischer, Aliza & Magen, Nathan, 2016. "Trust and reputation in the sharing economy: The role of personal photos in Airbnb," Tourism Management, Elsevier, vol. 55(C), pages 62-73.
- Fogg, Andrew, 2016. “Anthony Goldbloom gives you the secret to winning Kaggle competitions”, website: https://www.import.io/post/how-to-win-a-kaggle-competition/, downloaded on: 2020.06.11.
- Friedman, Jerome H., 2002. "Stochastic gradient boosting," Computational Statistics & Data Analysis, Elsevier, vol. 38(4), pages 367-378, February.
- Heckman, James J. & Golsteyn, Bart H.H. & Borghans, Lex & Meijers, Huub 2009. "Gender Differences in Risk Aversion and Ambiguity Aversion," NBER Working Papers 14713, National Bureau of Economic Research, Inc.
- Inside Airbnb, 2020a. “Get the Data”, website: http://insideAirbnb.com/get-the-data.html, downloaded on: 2020.02.27.
- Inside Airbnb, 2020b. “About”, website:http://insideAirbnb.com/about.html, downloaded on: 2020.02.27.
- Kakar, Venoo & Voelz, Joel & Wu, Julia & Franco, Julisa, 2018. "The Visible Host: Does race guide Airbnb rental rates in San Francisco?", Journal of Housing Economics, Elsevier, vol. 40(C), pages 25-40.
- Knaus, Michael C., 2018. "A Double Machine Learning Approach to Estimate the Effects of Musical Practice on Student's Skills," IZA Discussion Papers 11547, Institute of Labor Economics (IZA)

- Kyung, Minjung & Gill, Jeff & Ghosh, Malay & Casella, George, 2010. “Penalized Regression, Standard Errors, and Bayesian Lassos”, Bayesian Analysis. 5. 369-412. 10.1214/10-BA607.
- Marchenko, Anya, 2019. "The impact of host race and gender on prices on Airbnb," Journal of Housing Economics, Elsevier, vol. 46(C)
- Mason, Llew & Baxter, Jonathan & Bartlett, Peter & Frean, Marcus, 2002. “Boosting Algorithms as Gradient Descent”, Advances in Neural Information Processing Systems 12.
- Nielsen, Didrik, 2016. “Tree Boosting with XGBoost”, Master’s Thesis at the Norwegian University of Science and Technology, Department of Mathematical Sciences
- Oskam, Jeroen & Rest, Jean-Pierre & Telkamp, Benjamin, 2018. "What’s mine is yours—but at what price? Dynamic pricing behavior as an indicator of Airbnb host professionalization," Journal of Revenue and Pricing Management, Palgrave Macmillan, vol. 17(5), pages 311-328, October
- Perez, Saeta Israel, & Elmas, Ferhat, 2016. “gender-guesser Project description”, website: https://pypi.org/project/gender-guesser/, downloaded on: 2020.02.27.
- Perez-Sanchez, V. Raul & Serrano-Estrada, Leticia & Marti, Pablo & Mora-Garcia, Raul-Tomas, 2018. "The What, Where, and Why of Airbnb Price Determinants," Sustainability, MDPI, Open Access Journal, vol. 10(12), pages 1-31, December
- Stone, Mervyn, 1974. “Cross-Validatory Choice and Assessment of Statistical Predictions”. Journal of the Royal Statistical Society, Series B (Methodological), Volume 36, No. 2, pages 111-147.
- Walters, Amy E. & Stuhlmacher, Alice F. & Meyer, Lia L., 1998. "Gender and Negotiator Competitiveness: A Meta-analysis," Organizational Behavior and Human Decision Processes, Elsevier, vol. 76(1), pages 1-29, October
- World Health Organization, 2020. “WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020”, website:

https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020, downloaded on: 2020.06.11.

- xgboost developers, 2020. “XGBoost Python Package”, website: https://xgboost.readthedocs.io/en/latest/python/, downloaded on: 2020.05.16.
- Yang, Jui-Chung & Chuang, Hui-Ching & Kuan, Chung-Ming, 2020. "Double machine learning with gradient boosting and its application to the Big N audit quality effect," Journal of Econometrics, Elsevier, vol. 216(1), pages 268-283
- Zhang, Zhihua & J. C. Chen, Rachel & Han, Lee D. & Yang, Lu, 2017. "Key Factors Affecting the Price of Airbnb Listings: A Geographically Weighted Approach," Sustainability, MDPI, Open Access Journal, vol. 9(9), pages 1-13, September

Data availability

All data used in this thesis was downloaded from the Inside Airbnb website (http://insideairbnb.com/get-the-data.html). Detailed individual results and the codes used to perform the analysis are available at:

https://drive.google.com/drive/folders/1eCKHAGyIv76Lx5c_iE64Pws4fAaD7rU9?usp=sharing