Transport Policy 60 (2017) 10–23

Contents lists available at ScienceDirect

Transport Policy

journal homepage: www.elsevier.com/locate/tranpol

Identifying odometer fraud in used car market data☆

Josef Montag

International School of Economics, Kazakh-British Technical University, Tole bi st. 59, 050000 Almaty, Kazakhstan

ARTICLE INFO

JEL classification:
K42
R40

Keywords:
Used car market
Odometer fraud
Digit tests

ABSTRACT

This paper investigates the presence of odometer fraud in the used car market using a large dataset of car sale advertisements from the Czech Republic. The strategic aspects of sale decisions and the practice of rounding odometer readings, however, render the standard statistical tests for fabricated data invalid. I therefore develop and employ a modification of the last-digit test, which has been used to detect fraud in election data. Simulations using the data from advertisements and travel survey data from the United States support the validity of this test under alternative distributional assumptions. The results suggest that suspicious patterns are more prevalent in the segment of cars imported from abroad. I also show that this test can be used at the seller level, which may be of interest to authorities and market participants.

1. Introduction

It is commonly believed that odometer fraud plagues the used car market. However, it is difficult to detect.[1] Searching the phrase "mileage correction service" on Google yields 23.7 million results, while odometer correction tools can be bought for about $300.[2] In the European Union, about five percent of consumers who had purchased a second hand car stated that they experienced odometer fraud (European Commission, 2014). In Poland, the largest importer of used cars within the EU, the figure reaches 15 percent. An earlier study by the U.S. Department of Transportation (2002) estimated that 3.5 percent of all automobiles had their odometer rolled back, implying 450,000 cases of odometer fraud in the United States per year. The direct costs to US customers were estimated at one billion dollars.

However, scientific knowledge on the phenomenon of odometer fraud and its impact on the used car market is lacking. This paper takes the first steps to closing this gap and contributes in the following ways: (i) it explores the statistical techniques for fraud detection and ascertains the main challenges to their validity in the used car market setting; (ii) it develops an appropriate statistical technique, a modified version of the last-digit test that has been previously used to detect election fraud (Beber and Scacco, 2012), and documents its validity for detection of odometer fraud in the used car market setting; (iii) it tests for the presence of odometer fraud using a unique dataset on car sale advertisements from the Czech Republic, pinpointing the market segments with higher prevalence of suspected fraud, and uses the technique to identify potentially suspicious sellers; and (iv) it points to an economically important agenda open for future research.

Specifically, the paper looks at how the digit-based statistical tests can be used to identify the presence of odometer fraud in the used car market, in its segments, and sellers' offers. These techniques have been established as tools for the detection of accounting fraud, electoral fraud, or fabricated data from clinical trials (Al-Marzouki et al., 2005; Beber and Scacco, 2012; Debreceny and Gray, 2010; Klimek et al., 2012; Mebane, 2010).[3] Digit tests exploit the simple idea that true data, be it vote counts or sales books, are generated by random processes while fraudulent data are produced by people, and people are not good randomizers. This leads to statistical tests that allow us to distinguish true records from potentially fabricated data. One limitation of digit tests is that they cannot be used to identify fraud at the level of individual data entries, such as ballot box count or in our context an individual odometer reading. Notwithstanding this qualification these methods can identify suspicious patterns at more aggregate levels, such as a market segment or a used car dealer, and may thus provide valuable signals for policy makers, law enforcers, and market participants.

☆ I thank Peter Bolcha, Pavel Čížek, Petr David, Máté Fodor, Marek Litzman, Aslan Tanekenov, three anonymous reviewers, and participants at the 2015 Conference of the European Association of Law and Economics at the University of Vienna for helpful discussions. Any remaining errors should be attributed only to me. This research was funded by the Education for Competitiveness Operational Programme, project no. CZ.1.07/2.3.00/30.0031, co-financed by the European Social Fund and the governemnt budget of the Czech Republic.

E-mail address: josef.montag@gmail.com.

[1] See "Information about odometer fraud," the U.S. Department of Transportation, Online, February 2010, at https://www.nhtsa.gov/staticfiles/nvs/pdf/811284.pdf (last accessed on May 16, 2017).

[2] See offers on eBay.com at http://www.ebay.com/bhp/odometer-correction (last accessed on May 16, 2017).

[3] See also Benford (1938); Cho and Gaines (2007); Diekmann (2007); Jacob and Levitt (2003); Judge and Schechter (2009); Nye and Moul (2007); Rauch et al. (2011); Varian (1972).

http://dx.doi.org/10.1016/j.tranpol.2017.07.018
Received 27 December 2016; Received in revised form 10 July 2017; Accepted 29 July 2017

Two simple and tried techniques for the statistical detection of fraud are first-digit tests and last-digit tests. The first digit-approach exploits the fact that the distribution of the first digits in randomly generated count data generally follow the Benford distribution. The last-digit approach exploits the fact that the last digits of almost any count data are draws from the uniform distribution.

However, two important features of the used car market present a challenge for detecting odometer fraud with the help of these statistical methods. First, an individual's decision to sell a car is not random. For instance, the manufacturer warranty may be limited by mileage, possibly creating a discontinuity in the car's value at the warranty threshold and thus a discontinuity in the incentives to sell the car. This makes the use of the first-digit test problematic, since it has been shown to be invalid in situations where the data is a result of strategic decisions, typically elections (Deckert et al., 2011; Mebane, 2011). This concern also plausibly extends to the market for used cars. The last-digit test has been shown to be more robust as it remains valid under a wide range of assumptions about the data generating process (see Beber and Scacco, 2012).

The second challenge is posed by the common practice of rounding odometer readings in listed advertisements (ads), which may be strategic, yet innocuous. To further complicate the matters, the rounding is not consistent across advertisements. Some ads apparently list the full odometer readings, but readings rounded to the nearest ten, hundred, or thousand are also common. As a consequence, the distribution of the last digits of the advertised mileage would deviate from the uniform distribution even in the absence of fraud. This limits the direct usability of the last-digit tests for odometer fraud detection.

In this paper I therefore develop and employ a simple modification of the last-digit test and document its validity. The idea is that the digit structure of the mileage declared in an advertisement can be used to infer whether a specific digit has been affected by rounding or not. Intuitively, consider any integer whose last digit differs from zero. For such a number, one can infer that the last digit of this number has not been rounded and, crucially, that none of the preceding (higher order of magnitude) digits has been affected by rounding. This allows the identification of ads with odometer readings that have not been rounded. For this subset of data, the last-digit test can be employed as a test for fraud by testing the uniformity of the distribution of the penultimate and higher-order digits. Monte Carlo simulations using Czech data on used cars for sale and the National Household Travel Survey data from the United States support the validity of this approach under alternative distributional assumptions.

2. Materials and methods

2.1. Data

This paper uses a unique dataset provided by a company maintaining one of the main Czech web portals for advertising the sale of used cars.[4] The website allows firms and individuals to post their offers and the site appears on the first page of the Google. cz search for “used cars.”[5] Among the websites that state the total number of active advertisements, this site is one of the largest. The fee for a standard ad is about CZK 550 (EUR 20) and the fee for an ad featuring a photo is CZK 2000 (EUR 74). I did not find anything that would make the website distinct (from the sellers' or buyers’ viewpoint) from other similar services.

The provided database covers the full universe of advertisements posted on this website during the 22-month period between January 1, 2012 and November 5, 2013. Complete documentation was provided together with the database. The raw dataset of advertisements contains 440,052 records with 38 columns. Key entries include an identifier for the seller, the dates when the ad was placed and when it ended, the car's mileage, production year, whether the vehicle's service history is available, the country of origin, and the Vehicle Identification Number (VIN), if provided by the seller. The data, however, does not contain any information as to whether the seller is a private person or a company, nor does it contain a unique identifier for each ad.

Individual ads are placed on the website for a default period of 30 days, and can be renewed thereafter. As a result, an ad may occur in multiple records, depending on whether it was placed only once or whether it was subsequently renewed. Because ads and vehicles are not uniquely identified in the data (sellers need not supply the VIN code), identifying duplicated ads posed a major challenge, the technical details of which are provided in Appendix A. In summary, duplicated ads were identified using two approaches: (i) for ads in which a VIN code was supplied, I designed a program that tests the structure and content of each VIN code in order to separate valid VIN codes from those invalid and to check whether the vehicle characteristics match the stated VIN; (ii) among the ads that did not contain a valid VIN code, duplicated sale offers were identified using the subset of ads with valid VIN codes to obtain a set of variables with identical values across duplicated records in 95 percent of cases, which resulted in 12 variables altogether. These 12 variables were then used to identify duplicated records within ads placed by each seller in the rest of the data.

Because some of the ads on the website are for new automobiles or cars used by sellers as show cars, I restrict the database to automobiles between 2 and 30 years of age with mileage ranging from 1000 to 400,000 km. For brevity, I drop ads in which sellers do not state the vehicle's country of origin so that it is not clear whether the car was imported or not. However, the results for this subset of automobiles are qualitatively similar to the results for imported cars and are available upon request. The resulting analysis dataset contains 135,295 unique ads with a valid VIN code and 127,192 records identified as unique for each seller across the 12 variables, i.e. 262,487 observations altogether.[6]

2.2. Descriptive statistics

The summary statistics of the analysis dataset are reported in Table 1. The average car in the dataset has the stated mileage of almost 140,000 km and was produced in the middle of 2004. Over two thirds of ads state that the car's service history is available, and over one half of sellers provide a valid VIN code. Czech-produced Skoda vehicles are the most often offered make, followed by Ford and Volkswagen. Slightly more than one half of the ads are for cars of Czech origin, i.e. cars registered in the Czech Republic under their last owner, while more than one fourth of the cars are imported from Germany, and about 20 percent are imported from other countries.

The bottom section of Table 1 reports descriptive statistics for the last five digits of the odometer readings. The means of distributions for the last, the penultimate, and the third-last digits are well below the expected value of 4.5. However, as discussed below in more detail, this is largely an artifact of rounding. To further illustrate possible irregularities in the declared odometer readings, histograms of odometer readings are plotted in Fig. 1. The data are split by service history and VIN code availability, the place of origin, and fuel type, and are binned into 5000 km categories. In general, the heights of any two neighboring bins should be similar. However, this is often not the case in the data as some bins appear to be more “popular” than their neighbors. These apparent irregularities would probably not occur in data with true odometer readings from a random sample of the car population. However, these patterns might also be explained by the strategic nature of rounding, or decision to sell a car, i.e.

[4] The company that provided the database wishes to remain anonymous. However, contact with them can be facilitated upon request.

[5] See https://www.google.cz/#q=ojeta+auta (last accessed May 16, 2017).

[6] The analysis dataset and code replicating the results reported in this paper are available at https://sites.google.com/site/josefmontag/josef/research or upon request. Statistical analysis was performed in R 3.3.3 (R Core Team, 2017).