

An Empirical Investigation of Biased Survey Data

James E. Prieger
Department of Economics
University of California
One Shields Avenue
Davis, CA 95616-8578
telephone: (530) 752-8727
fax: (530) 752-9382
email: jeprieger@ucdavis.edu

January 11, 2005

Abstract

This paper investigates response bias in survey data on annual driving mileage and evaluates the performance of an econometric remedy proposed in the literature, Orbit. There are three contributions in this paper. I first discuss pseudo-precision bias, caused by asking respondents to quantify something they have never precisely measured, and estimate the bias in a self-report of annual mileage driven. The estimation of the bias accounts for the non-standard censoring of the data. Individuals systematically exaggerate their deviation from the sample average, and using the self-reported data leads to misleading estimates of the income elasticity of travel mileage. Second, I extend the Orbit procedure, which has been proposed to correct for reporting bias, to allow misreporting at the lower censoring point. This generalized version may be useful in other settings as well. Finally, I demonstrate that in this application Orbit does not improve the accuracy of the estimation and does not correctly uncover the direction of the pseudo-precision bias. The message for practitioners using biased data is therefore a cautionary one.

Keywords: Orbit, response bias, travel demand, semiparametric estimation, censoring, ordered choice data.

1 Introduction

Many microeconomic studies make use of survey data, such as answers to questions on an individual's job history, consumer expenditure, and use of technology. A potential concern with any self-reported information is response bias. In this paper I investigate response bias in answers to a particular survey question—how many miles the respondent drove last year—and evaluate the performance of a remedy proposed in the econometric literature. I find that there is systematic response bias, with individuals exaggerating their deviation from the sample average, and that using the self-reported data leads to misleading estimates of the income elasticity of miles driven. Klein and Sherman's (1997) Orbit estimator, which is designed to correct for reporting bias, fails to detect the nature of the bias and distorts the income elasticity estimate even further.

I make three contributions in this paper. First, I discuss *pseudo-precision bias*, caused by asking respondents to quantify something they have never precisely measured, and estimate the bias in a self-report of annual mileage driven. This estimation is non-trivial its own right due to non-standard censoring of the data. I find that using the misreported data to estimate the determinants of mileage leads to substantial bias in the estimates. Second, I extend Klein and Sherman's (1997) Orbit estimator to match the requirements of the present application. This generalized version may be useful in other settings as well. Finally, I demonstrate that Orbit estimation using the self-reported data does not improve the accuracy of the estimated coefficients, and furthermore does not correctly uncover the direction of the pseudo-precision bias either. Thus, in this particular application no cure is found for the problems caused by response bias. The message for practitioners using survey data where pseudo-precision bias is a concern is therefore a cautionary one. As promising as techniques like Orbit are for undoing the ills caused by biased self-reported measurements, in some applications the assumptions required for consistent estimation will fail and there is no substitute for accurate data. Without more accurate data one will not even be able to assess how badly Orbit is failing.

I begin in section 2 by investigating the relationship between self-reported and odometer-measured mileage using observations on both. I then explore the effects of various covariates on miles driven in section 3, with an emphasis on estimating the income elasticity. The income elasticity of mileage for low-income households is estimated to be much higher when self-reported miles are used than when measured miles are used. The Orbit estimate of the income elasticity is even higher yet. Furthermore, I show in section 4 that the Orbit estimate of the response bias function fails to reveal the marked exaggeration by drivers who claim very low or very high mileage.

2 Investigating the Response Bias

To correctly answer a measurement question on a survey, the information must be accessible to a respondent. Accessibility requires that the respondent had the information at some point and understood it, and that he can retrieve it from memory (Kalton and Schuman, 1982). In most surveys there is no way to directly measure response bias due to inaccessibility, because there is no objective, external measurement of the quantity in question. Rarely authors in the economics literature have been able to link to external data to verify the accuracy of the survey responses (see Oyer (2004) and citations therein). These studies focus on recall bias of job history. Recall bias pertains to the memory retrieval part of accessibility. In contrast, in this study I examine the other component of inaccessibility bias, which I term *pseudo-precision bias*: that due to asking for factual information the respondent never possessed in the first place.

To illustrate these two components of inaccessibility bias, consider a survey question asking a part-time worker how many hours she worked last year. Pseudo-precision bias come from her estimating the answer, if she never calculated her hours at year end. Recall bias comes into play if she did the calculation in the past and has to remember the figure. There may be many daily activities in which survey respondents do not normally quantify their amount of participation. Examples of survey questions about these less-quantified activities include asking how often individuals use

a cellular telephone while driving (Hahn and Prieger, 2004), how many days a home computer is used each week (the Computer Use and Ownership supplement to the 1997 Current Population Survey), and how much is spent on average gambling each month (Lesieur, 1998). Errors in such reports are not caused by recall bias, but instead from the respondent's attempt to quantify for the sake of the survey that which he may never have quantified before in his daily life for his own purposes. Often researchers using such survey data take them at face value, despite the possibility of false precision.

The 2001 National Household Travel Survey (NHTS)¹ by the U.S. Department of Transportation offers an interesting opportunity to explore pseudo-precision bias in such measurement questions. The survey asks drivers how many miles they drove their vehicle in the past 12 months. I call the answer to this question *SRMILES*, for self reported mileage. Odometer readings were also taken by the respondent at two points in time, the first of which was usually at the time of the main survey questioning. From the odometer readings, the survey administrator created an annualized estimate of miles driven by the vehicle during the period May 2001–April 2002. The estimate adjusts for the duration between odometer readings and seasonal and annual differences in miles traveled. I call this measurement-based mileage variable *ODOMILES*. For the most accurate results, I restricted the sample to observations where *SRMILES* and *ODOMILES* should be close to each other if *SRMILES* were accurately reported.² Discrepancies between the two variables are thus due to two factors: errors caused by the annualization of the odometer readings, and reporting bias by the respondent.

The data are plotted in Figure 1. One unusual feature of the NHTS data is the non-standard censoring of *ODOMILES*, which is immediately apparent in Figure 1. When the ratio of *SRMILES* to *ODOMILES* is greater than four or less than one-quarter, and the absolute deviation is greater

¹See <http://nhts.ornl.gov/2001/index.shtml> for documentation.

²In particular, the sample is restricted to observations where the respondent is answering questions about his or her own driving and is the primary driver of a single vehicle. Motorcycles and RVs were excluded, as were vehicles with odometer readings taken less than 90 days apart. Finally, I required a minimum of six months overlap between the coverage period for *ODOMILES* (May 2001–April 2002) and the period covered by *SRMILES*.

than 10,000 miles, *ODOMILES* is not reported. About 8.5% of the odometer readings in the estimation sample are censored above or below.

Do respondents accurately report mileage? On average, yes. The means of odometer-measured and self-reported miles (for the uncensored observations) are 11,912 and 11,728, respectively. However, the averages mask the imprecision in the relationship: correlation is only 0.64 between the uncensored observations. It would most likely be lower if the censored *ODOMILES* were able to be observed, because it is the observations for which the two variables differ widely that are censored. To further explore the relationship between *SRMILES* and *ODOMILES*, I use piecewise polynomial splines and nonparametric smoothing. Mileage is measured in logs in all estimations, although depicted in levels on the graphs. The censoring precludes the use of simple OLS for consistent estimation of the piecewise polynomial splines, for reasons familiar to tobit model settings. The parametric splines are fitted by MLE, assuming normal errors in a log-log specification of *SRMILES* on *ODOMILES* (see the appendix for the likelihood). The resulting fit, depicted as the heavy line in Figure 1, shows that odometer-measured mileage exceeds self-reported mileage up to about 5,000 miles, is roughly equal to self-reported mileage between 5,000 and 13,000 miles, and is less than self-reported mileage for higher mileage. Friedman's (1984) variable-span local averaging smoother, drawn in Figure 1 with the lighter solid line, shows a similar relationship between the two measures of mileage. The smoother does not account for the censoring, however, and the discrepancy up to 5,000 miles does not look as severe because the above-censored observations in that region are dropped.

I interpret the discrepancy between fitted odometer-measured mileage and self-reported mileage as pseudo-precision bias. In this view, then, drivers underreport actual mileage up to 5,000 miles and overreport actual mileage beyond 13,000 miles. In other words, drivers exaggerate their deviation from the average when they are far from it. The other potential cause of the discrepancy is measurement error in *ODOMILES* caused by the annualization of the odometer readings. However, this measurement error should be mean zero by construction because the annualization uses

national averages to account for seasonal and yearly differences in mileage. Censoring complications aside, mean-zero measurement error in the dependent variable does not lead to inconsistent estimates (as long as it is uncorrelated with the regressors). Recall bias is less germane with this question, unless the driver calculated mileage weekly or monthly during the course of the past year, which is unlikely for most. Therefore it appears safe to interpret systematic differences between *SRMILES* and *ODOMILES* as pseudo-precision bias.

3 Estimating the Income Elasticity of Mileage

Now that the form of the reporting bias is known, I turn to estimating the determinants of miles driven. For the sake of concreteness, here I assume the relationship of interest is the income elasticity of annual mileage, although the same conclusions would be reached with many of the other variables as well. The income elasticity may be of interest, for example, to a regional planner trying to forecast transportation infrastructure needs over a forecast period in which household incomes are expected to grow. A second example is the problem of evaluating the effects of environmental legislation on gasoline consumption and pollution, in which the income elasticity of mileage may play a role (Goldberg, 1998). I first estimate the income elasticity using the odometer-measured mileage, which is taken to be the best estimate of the true income elasticity. Economists often lack objective measures of quantities of interest, however, and must instead use survey data to estimate relationships between variables. I next demonstrate that using the self-reported mileage instead of the odometer-measured mileage leads to poor estimates of the income elasticity. Furthermore, an *a priori* promising method to correct for survey bias, Orbit, does not improve the estimate but instead exacerbates the bias.

In addition to income, many other factors such as age, gender, employment status, and mode of travel to work may affect annual mileage. I control for these variables in all estimations, as well as for vehicle type and age, rural location, and Census region. The estimation results when the

dependent variable is *ODOMILES* in logs are in the first columns of Table 1. The income elasticity significantly differs from zero only for the lowest household income quartile group (*Income1*), for which it is a relatively inelastic 0.14. Thus the income elasticity of annual mileage with respect to income is very small, and significant only for lower-income households. This finding accords with the small income elasticities of gasoline demand found by Goldberg (1998) and others. Of the other significant coefficients, newer vehicles, vans and minivans, male drivers, employment, and rural location are all correlated with higher annual mileage. Age shows an inverted-U pattern with peak mileage at age 35.

In a typical survey setting, a variable such as *ODOMILES* would not be available to the researcher, and estimation would instead be based on self-reported miles. The same estimation repeated using *SRMILES* as the dependent variable is in the middle columns of Table 1. There is no censoring of *SRMILES* and OLS can be used to estimate the income elasticity. The elasticity estimate is again significant at the 5% level only for the first income quartile group. However, the estimate *Income1* is 0.23, nearly twice as high as (and outside the 95% confidence interval of) the previous estimate. The same pattern applies to the other coefficients in the specification as well: the significance of each is typically similar to the first estimation, but the magnitudes are often quite different. Thus, in the examples mentioned above, the self-reported measure could lead to misleading forecasts of demand for transportation infrastructure such as roadways, or to incorrect predictions of the effects of environmental policy on gasoline consumption.

If one suspects that miles are measured with error when self-reported and other data are not available, then one may look to the econometric literature for estimators designed to be consistent in the presence of reporting bias of unknown form. One of the few such estimators appropriate for continuous data is Klein and Sherman's (1997) (hereafter KS) Orbit estimator.³ The model for the

³Other estimators available in the literature for biased survey data (Klein and Sherman, 2002; Hsiao and Sun, 1999) are designed primarily for discrete data.

true quantity Q and the reported quantity Y is assumed to be

$$Q = \Lambda(Y) = (X'\beta_0 + u)^{1-d} c^d \quad (1)$$

$$d = 1\{X'\beta_0 + u \leq c\} \quad (2)$$

where $1\{\cdot\}$ is the indicator function, X and β_0 are k -vectors, u has a known parametric distribution, c is a censoring point, d is a censoring indicator, and Λ is the inverse of a strictly monotone reporting function mapping the truth into the report. Three advantages of the Orbit estimator are that it places few restrictions on Λ other than monotonicity (for example, Λ need not be differentiable), it can handle censored data, and it is relatively easy and quick to implement, unlike many semiparametric estimators. One potential disadvantage in some applications is that Λ must have two fixed points at known values: $\Lambda(c) = c$ and $\Lambda(s) = s$ for some $s > c$. Without this latter assumption Λ is identified only up to location and scale. Assume that u is distributed $\mathcal{N}(0, \sigma_0^2)$. Define $z = (y, x) \in [c, \infty) \times \mathbb{R}^k$, $\theta = (\beta, \sigma^2) \in \Theta$, a compact subset of $\mathbb{R}^k \times \mathbb{R}_{++}$. For each $\lambda_1, \lambda_2, t_1, t_2$ (each greater than c) define

$$\begin{aligned} f(\lambda_1, \lambda_2, t_1, t_2, z, \theta) &= 1\{y \leq t_1\} \log \Phi\left(\frac{\lambda_1 - x'\beta}{\sigma}\right) \\ &+ 1\{t_1 < y \leq t_2\} \log \left[\Phi\left(\frac{\lambda_2 - x'\beta}{\sigma}\right) - \Phi\left(\frac{\lambda_1 - x'\beta}{\sigma}\right) \right] \\ &+ 1\{y > t_2\} \log \left[1 - \Phi\left(\frac{\lambda_2 - x'\beta}{\sigma}\right) \right] \end{aligned} \quad (3)$$

where Φ is the standard normal cdf. If u is not normally distributed, then Φ in (3) is replaced with the cdf of u . The Orbit estimator of θ proposed by KS given a sample of independent observations from model (1), z_1, \dots, z_N , is

$$\hat{\theta}(s) = \operatorname{argmax}_{\theta} \frac{1}{N} \sum_{i=1}^N f(c, s, c, s, z_i, \theta) \quad (4)$$

Orbit is so-named because $\hat{\theta}$ is a maximum likelihood estimator for ordered choice data (along with the tobit form of (1)). As such, it is \sqrt{N} -consistent for θ_0 and asymptotically normal, and the asymptotic standard errors reported by any econometric software for MLE will be correct.

The promise of the Orbit estimator is clear: if (1) is correctly specified, then even if respondents misreport the truth almost everywhere across the range of the data the true parameter value θ_0 can be consistently estimated. No parametric or continuity assumptions are required of Λ , beyond monotonicity and known value at two points. If the parametric assumption on u in (1) is of concern, Klein and Sherman (2002) extend the procedure to allow estimation of θ_0 without specifying a distribution for u .

Note in passing that the reporting bias cannot be treated as classical measurement error in the dependent variable.⁴ Classical measurement error in the dependent variable in a linear regression model lowers the precision of the estimates, but does not make them inconsistent. Set aside censoring complications for the moment. The nature of the reporting function here is to destroy the linear form of the right side of (1), so that $E(Y)$ is no longer a linear function of the regressors. Even if Λ is linear (with slope a), a regression of Y on X will not yield consistent estimates of β_0 ($\hat{\beta}$ has plim β_0/a in this case).

For the present application, $c = 0$ (i.e., there can be no negative miles). Because the assumption that $\Lambda(0) = 0$ is problematic given that the data are to be transformed into logs (and furthermore given the overreporting demonstrated in the previous section for small values of $SRMILES$), I extend the Orbit estimator slightly. If the two fixed points of Λ required for identification are instead $s_1 > c$ and $s_2 > s_1$ satisfying $\Lambda(s_1) = s_1$ and $\Lambda(s_2) = s_2$, then let the extended Orbit estimator for θ_0 be

$$\hat{\theta}_2(s_1, s_2) = \operatorname{argmax}_{\theta} \frac{1}{N} \sum_{i=1}^N f(s_1, s_2, s_1, s_2, z_i, \theta) \quad (5)$$

As with Orbit, the asymptotic standard errors reported by any econometric software for MLE for $\hat{\theta}_2$ will be correct.

This version of Orbit is useful in any application where the data may be misreported at the censoring point. The results from this estimation, with $s_1 = 5,000$ and $s_2 = 8,000$ (chosen based on

⁴By classical measurement error I mean that $Y = Q + \varepsilon$, where ε is an error uncorrelated with X .

Figure 1), are in the final columns of Table 1.⁵ The elasticity estimate is again strongly significant only for the first income quartile group. The estimate *Income1* is 0.35, even higher than (and outside the 95% confidence interval for) the OLS estimate using the self-reported data and 2.5 times as high as the estimate using *ODOMILES*. Thus, in this application, not only does Orbit fail to correct the bias evidenced in the OLS estimate using *SRMILES*, it moves the estimate in the wrong direction.⁶

The failure of the estimator in this application is not due to any intrinsic flaw of the extended Orbit procedure. Table 2 contains the results of a Monte Carlo exercise carried out when the data are generated from the model postulated for Orbit, equations (1) and (2), with the piecewise linear reporting function:

$$\Lambda^{-1}(Q) = [0.5(Q + 1)]^{1\{Q < 2\}} [1.5(Q - 1)]^{1\{Q \geq 2\}} \quad (6)$$

which has fixed points at 1 and 3. Data for the two regressors (x_1, x_2) were generated from an equal-probability mixture of a mean-zero bivariate normal distribution with unit variances and covariance of 0.5, and a bivariate distribution with the unit-rate exponential distribution for x_1 and the $\chi^2(1)$ distribution for x_2 , where there is no correlation between the exponential and chi-square distributions. This mixture distribution generates data (x_1, x_2) that is correlated and highly skewed, just as the actual data in our application is. Q was censored at zero. Table 2 shows that the coefficient estimates are highly accurate on average, with bias on the order of 0.001 or smaller.

⁵It is worth noting that without the information from Figure 1, which the researcher would not have if *ODOMILES* were not available, selection of s_1 and s_2 would be more difficult (KS provide guidance on selection of the fixed points, however).

⁶If the fixed points s_1 and s_2 are incorrectly chosen but the model is otherwise correctly specified, then $\hat{\theta}$ will be biased but $\hat{\beta}/\hat{\sigma}$ is still consistent for β_0/σ_0 for the slope elements of β_0 . However, even if the ratio of the Orbit coefficient for *Income1* to $\hat{\sigma}$ is examined instead of the coefficient itself, it is still more than twice the same ratio from the estimates using *ODOMILES*.

4 Estimating the Response Bias with Orbit

The Orbit procedure also allows semiparametric estimation of Λ itself. The next question, then, is whether Orbit can correctly uncover the response bias pattern shown in Figure 1, despite its failure to accurately estimate β_0 . Let \mathcal{L}_t be a compact subset of \mathbb{R} containing $\Lambda(t)$ for each $t > c$. Define

$$f^-(\lambda, t, s, z, \theta) = f(\lambda, s, t, s, z, \theta) \quad (7)$$

$$f^+(\lambda, t, s, z, \theta) = f(s, \lambda, s, t, z, \theta) \quad (8)$$

where f is as defined in (3). The Orbit estimator for Λ at t is

$$\hat{\Lambda}(t; s) = \operatorname{argmax}_{\lambda \in \mathcal{L}_t} \frac{1}{N} \sum_{i=1}^N f^-(\lambda, t, c, z_i, \hat{\theta}(s)) \quad (9)$$

Note that the estimate of θ_0 from the first stage is used in this second stage. My extended estimator for Λ is

$$\hat{\Lambda}^-(t, s_2; s_1) = \operatorname{argmax}_{\lambda \in \mathcal{L}_t} \frac{1}{N} \sum_{i=1}^N f^-(\lambda, t, s_2, z_i, \hat{\theta}_2(s_1, s_2)) \text{ for } t \in [y_{\min}, s_2] \quad (10)$$

$$\hat{\Lambda}^+(t, s_1; s_2) = \operatorname{argmax}_{\lambda \in \mathcal{L}_t} \frac{1}{N} \sum_{i=1}^N f^+(\lambda, t, s_1, z_i, \hat{\theta}_2(s_1, s_2)) \text{ for } t \in [s_1, \infty) \quad (11)$$

where y_{\min} is the minimum y in the data. Note that for $t \in [s_1, s_2]$ two estimates are available and must be reconciled in some way. In this application (and in the Monte Carlo exercises I performed) the difference between estimates $\hat{\Lambda}^-$ and $\hat{\Lambda}^+$ for $t \in (s_1, s_2)$ is negligible (on the order of 0.02% at most) and I simply average the two. Thus the extended estimator is

$$\hat{\Lambda}_2(t; s_1, s_2) = \begin{cases} \hat{\Lambda}^-(t, s_2; s_1) & t \in [c, s_1] \\ \frac{1}{2}\hat{\Lambda}^-(t, s_2; s_1) + \frac{1}{2}\hat{\Lambda}^+(t, s_1; s_2) & \text{for } t \in (s_1, s_2] \\ \hat{\Lambda}^+(t, s_1; s_2) & t \in (s_2, \infty) \end{cases} \quad (12)$$

$\hat{\Lambda}_2$ is a step function with jumps at the sample values y_i . $\hat{\Lambda}(t)$ is \sqrt{N} -consistent for $\Lambda(t)$ and asymptotically normal and $\hat{\Lambda}$ converges in distribution to a Gaussian process; see KS for details. The extension of these results to $\hat{\Lambda}_2$ is straightforward. The pointwise variance of $\hat{\Lambda}_2(t; \cdot)$ cannot

be taken from the output from a standard MLE routine, due to the embedded estimate $\hat{\theta}_2$ in $\hat{\Lambda}_2$. Since $\hat{\Lambda}_2$ can be calculated quickly, bootstrapping is probably the easiest way to get asymptotically correct standard errors. The analytical form of the variance of $\hat{\Lambda}_2$ is in the appendix.

The estimate of Λ for the present application, with $s_1 = 5,000$ and $s_2 = 8,000$ as above, is shown in Figure 2.⁷ Comparing Figure 2 with the bias pattern revealed in Figure 1 shows that $\hat{\Lambda}_2$ does not uncover the actual reporting bias. $\hat{\Lambda}_2$ in Figure 2 suggests that self-reported miles mildly overreport *ODOMILES* up to 8,000 miles (best seen from the inset in the figure), after which gross underreporting sets in. However, this is exactly the opposite pattern from that found in Figure 1 where *ODOMILES* were actually used in the estimation.

Why does Orbit fail so badly to estimate the reporting bias? The problem lies not with the estimator itself. KS show that Orbit does well in simulations, which I verify for my extension of Orbit when the assumptions of the model are satisfied. Continuing with the Monte Carlo exercise discussed above, Figure 3 shows that the reporting function Λ^{-1} is estimated with high accuracy by $\hat{\Lambda}_2^{-1}$. The width of the 95% pointwise confidence interval shrinks to zero at the fixed points (1 and 3, in this example) by construction of $\hat{\Lambda}_2$. Away from the fixed points the confidence interval is still very narrow, although it begins to widen somewhat on the right, due to the skewed distribution of the regressors.⁸

The next suspicion, then, may be that the parametric assumption for u in (1) is incorrect. However, when u is assumed to be logistic, or when u is assumed to be normal but *SRMILES* and *ODOMILES* are in levels instead of logs, the same general bias pattern as in Figure 2 results. Neither varying s_1 and s_2 nor using the original version of Orbit instead of my extension changed the results qualitatively, either. Furthermore, normality appears to be a satisfactory assumption for u in (1), except possibly for the lower tail of the distribution. A QQ-plot (Figure 4) of the

⁷One of the great advantages of Orbit is its speed of implementation. Generating the data for the figure required estimating $\hat{\Lambda}_2$ at a grid of 300 t 's, each point of which requiring MLE. However, each MLE requires estimation of but a single parameter, and the whole process took only 75 seconds using the software package TSP on a PC with a 1.6Ghz processor.

⁸Less than 10% of the pseudo-data lie to the right of $Q = 4$ in Figure 3.

residuals from the regression of *ODOMILES* (from the first columns of Table 1) shows that, apart from the lower 1% of the tail, the residuals line up well with the standard normal distribution.⁹

A more likely reason for the failure of Orbit in this application is the nature of the misreporting itself. While the fitted lines in Figure 1 reveal a clear pattern of bias, the figure also shows that there is substantial deviation in the data from the average bias at any one value of *SRMILES*. The correlation between *SRMILES* and *ODOMILES* is only 0.65 for the uncensored observations (and it would probably be lower if completely uncensored observations were available), whereas if the data lined up perfectly with the fitted polynomial spline the correlation would be 0.92. The model in (1) assumes that the reporting function transforms the true measurement in a consistent, deterministic fashion. This may be true if the misreporting at issue is recall bias. For example, numerous studies have shown that individuals systematically bring events forward in time in their memory—a phenomenon known as *telescoping* (Kalton and Schuman, 1982). However, when pseudo-precision bias—the other component of inaccessibility bias—is at issue, (1) may not be a good model of the outcome of the cognitive processes of the respondent. Asking a respondent to report a measurement the individual has never taken may result in anything from random answers to informed “guesstimates”. In such cases the correct model would be considerably more complex than that in (1), and any assumed model would be untestable.

5 Conclusion

In this article I explore pseudo-precision bias, which is caused by asking survey respondents to quantify something they have never precisely measured, and an attempted cure. The driving mileage data I investigate exhibits systematic reporting bias, with respondents exaggerating their deviation from the mean in the tails of the distribution. The misreported data leads to substantial bias in

⁹In a QQ-(quantile-quantile) plot, if the distributional assumption for the regression error is correct the standardized residuals will be close to the 45-degree line when plotted against quantiles of the standard normal distribution. In the figure, residuals from censored observations were replaced with random draws from the appropriate truncated normal distribution (Dunn and Smyth, 1996).

the estimated income elasticity of mileage. An extension of Klein and Sherman’s (1997) Orbit estimator fails to improve the accuracy of the income elasticity estimate, and in fact increases the bias. Orbit also incorrectly estimates the direction of the pseudo-precision bias. Practitioners using survey data subject to pseudo-precision bias must therefore beware. No econometric technique can remove the need for more accurate data.

Although the assumptions necessary for Orbit apparently fail in this application, the extended version of Orbit I present may prove useful in other settings. For example, any model that is linear in the covariates in a region around the mean but deviates from linearity in the tails (subject to monotonicity of the transformation function) can be estimated with the extended version of Orbit.

A Appendix

A.1 Likelihood for Censored Mileage

Assume the model relating *ODOMILES* (y) to *SRMILES* (x) is

$$\log y = P_3(\log x) + \varepsilon$$

Here P_3 is a piecewise cubic spline function: $P_3(w) = \alpha + \sum_{i=1}^3 \sum_{j=0}^4 \beta_{ij}(w - \kappa_j)^i$, where (α, β) are parameters to be estimated, $\kappa_0 = 0$, and $\kappa_1, \dots, \kappa_4$ are the quintiles of *SRMILES*. The error ε is distributed $N(0, \tau_0^2)$. Let d be a censoring indicator that is zero if *ODOMILES* is not censored.

There are four types of censoring possible:

$$\begin{array}{llll} d = 1 & \text{if } 0 < x \leq 3,333.3 & \text{and } y > x + 10,000 & \text{(censored above)} \\ d = 2 & \text{if } 3,333.3 < x \leq 10,000 & \text{and } \frac{y}{x} > 4 & \text{(censored above)} \\ d = 3 & \text{if } 10,000 < x \leq 13,333.3 & \text{and } y < x - 10,000 \text{ or } \frac{y}{x} > 4 & \text{(censored above or below)} \\ d = 4 & \text{if } x > 13,333.3 & \text{and } \frac{y}{x} < \frac{1}{4} \text{ or } \frac{y}{x} > 4 & \text{(censored above or below)} \end{array}$$

When $d = 3$ or 4 , it is not known whether y is censored above or below. Define the following censoring functions:

$$c_1(x) = \begin{cases} -\infty & \text{if } d \leq 2 \\ \log(x - 10,000) & \text{if } d = 3 \\ \log(x) - \log(4) & \text{if } d = 4 \end{cases}$$

$$c_2(x) = \begin{cases} \log(x + 10,000) & \text{if } d = 1 \\ \log(x) + \log(4) & \text{if } d = 2 \\ \infty & \text{if } d \geq 3 \end{cases}$$

Then the log likelihood for an observation i is

$$\begin{aligned} \ln L_i &= 1\{d_i = 0\} \cdot \left\{ -\log(\sigma) + \log \left[\phi \left(\frac{\ln y_i - P_3(\log x_i)}{\sigma} \right) \right] \right\} \\ &\quad + 1\{d_i \geq 1\} \cdot \left\{ \log \left[\Phi \left(\frac{c_1 - P_3(\log x_i)}{\sigma} \right) + 1 - \Phi \left(\frac{c_2 - P_3(\log x_i)}{\sigma} \right) \right] \right\} \end{aligned}$$

A.2 The Variance of the Extended Orbit Estimator for Λ

First we find the pointwise variance of $\hat{\Lambda}^-(t)$. Assume that the data are an iid sample, that θ_0 is in the interior of Θ , that $\hat{\Lambda}^-(t)$ is in the interior of \mathcal{L}_t , and that the support of X is bounded. The latter assumption ensures that probabilities that are arguments of the log function are bounded away from zero. Then using methods analogous to KS, a second order Taylor expansion of the estimating equations for ML for $\hat{\theta}$ and $\hat{\Lambda}^-(t)$ leads to

$$\sqrt{N} \left(\hat{\Lambda}^-(t) - \Lambda(t) \right) = B^{-1} \sqrt{N} \left[E_N \nabla_{\lambda} f^-(\Lambda(t), \theta_0) + D A^{-1} E_N \nabla_{\theta} f(\theta_0) \right] + o_p(1) \quad (13)$$

where

$$A = -\text{plim } E_N \nabla_{\theta\theta} f(s_1, s_2, s_1, s_2, z_i, \theta_0) \quad (14)$$

$$B = -\text{plim } E_N \nabla_{\lambda\lambda} f^-(\Lambda(t), t, s_2, z_i, \theta_0) \quad (15)$$

$$D = \text{plim } E_N \nabla_{\lambda\theta'} f^-(\Lambda(t), t, s_2, z_i, \theta_0) \quad (16)$$

and E_N is the empirical measure $N^{-1} \sum_{i=1}^N$. From (13) it follows that

$$\sqrt{N} \left(\hat{\Lambda}^-(t) - \Lambda(t) \right) \xrightarrow{d} \mathcal{N}(0, V^-) \quad (17)$$

where

$$V^- = B^{-1} + DA^{-1}(2C + D')/B^2 \quad (18)$$

$$C = \text{plim } E_N \left[\nabla_{\theta} f(s_1, s_2, s_1, s_2, z_i, \theta_0) \nabla_{\lambda} f^-(\Lambda(t), t, s_2, z_i, \theta_0) \right] \quad (19)$$

The second term on the right side of (18) reflects the extra variance from the embedded estimate of $\hat{\theta}$ in $\hat{\Lambda}^-(t)$.

The argument for $\hat{\Lambda}^+$ proceeds similarly. Define H , J , and K for f^+ to be analogous to B , C , and D for f^- , respectively. Then the limiting variance of $\sqrt{N} \left(\hat{\Lambda}^+(t) - \Lambda(t) \right)$ is

$$V^+ = H^{-1} + KA^{-1}(2J + K')/H^2 \quad (20)$$

The remaining task is to find the limit distribution of $\hat{\Lambda}_2$ in the region $(s_1, s_2]$, which is

$$\sqrt{N} \left(\hat{\Lambda}_2(t) - \Lambda(t) \right) \xrightarrow{d} \mathcal{N}(0, V) \quad (21)$$

where

$$V = \frac{1}{4}(V^- + V^+) + \frac{1}{2}W \quad (22)$$

$$W = (P + C'A^{-1}K' + DA^{-1}M + DA^{-1}K') / (BH) \quad (23)$$

$$M = \text{plim } E_N \left[\nabla_{\theta} f(s_1, s_2, s_1, s_2, z_i, \theta_0) \nabla_{\lambda} f^+(\Lambda(t), t, s_1, z_i, \theta_0) \right] \quad (24)$$

$$P = \text{plim } E_N \left[\nabla_{\lambda} f^+(\Lambda(t), t, s_1, z_i, \theta_0) \nabla_{\lambda} f^-(\Lambda(t), t, s_2, z_i, \theta_0) \right] \quad (25)$$

The scalar W reflects covariance between $\hat{\Lambda}^-$ and $\hat{\Lambda}^+$. Feasible versions of A, B, C, D, H, J, K, M , and P can be found by taking sample averages (e.g., $A_N = -E_N \nabla_{\theta\theta} f(s_1, s_2, s_1, s_2, z_i, \hat{\theta})$ for A). As a practical matter, given the ease of calculating the Orbit estimator, the practitioner may want to find the variance of $\hat{\Lambda}_2$ by bootstrapping.

References

- Dunn, Peter K. and Smyth, Gordon K. (1996), ‘Randomized Quantile Residuals’, *Journal of Computational and Graphical Statistics* **5**, 236–244.
- Friedman, Jerome H. (1984), ‘A Variable Span Smoother’, Technical Report 5, Laboratory for Computational Statistics, Dept. of Statistics, Stanford University.
- Goldberg, Pinelopi Koujianou (1998), ‘The Effects of the Corporate Average Fuel Efficiency Standards in the US’, *Journal of Industrial Economics* **46**(1), 1–33.
- Hahn, Robert W. and Prieger, James E. (2004), ‘The Impact of Driver Cell Phone Use on Accidents’, Working Paper 04-14, AEI-Brooking Joint Center for Regulatory Studies.
- Hsiao, Cheng and Sun, Bao-Hong (1999), ‘Modeling Survey Response Bias—with an Analysis of the Demand for an Advanced Electronic Device’, *Journal of Econometrics* **89**(1–2), 15–39.
- Kalton, Graham and Schuman, Howard (1982), ‘The Effect of the Question on Survey Responses: A Review’, *Journal of the Royal Statistical Society. Series A (General)* **145**(1), 42–57.
- Klein, Roger and Sherman, Robert (1997), ‘Estimating New Product Demand from Biased Survey Data’, *Journal of Econometrics* **76**, 53–76.
- Klein, Roger W. and Sherman, Robert P. (2002), ‘Shift Restrictions and Semiparametric Estimation in Ordered Response Models’, *Econometrica* **70**(2), 663–691.
- Lesieur, Henry R. (1998), ‘Costs and Treatment of Pathological Gambling’, *Annals of the American Academy of Political and Social Science* **556**, 153–171.
- Oyer, Paul (2004), ‘Recall Bias Among Displaced Workers’, *Economics Letters* **82**, 397–402.

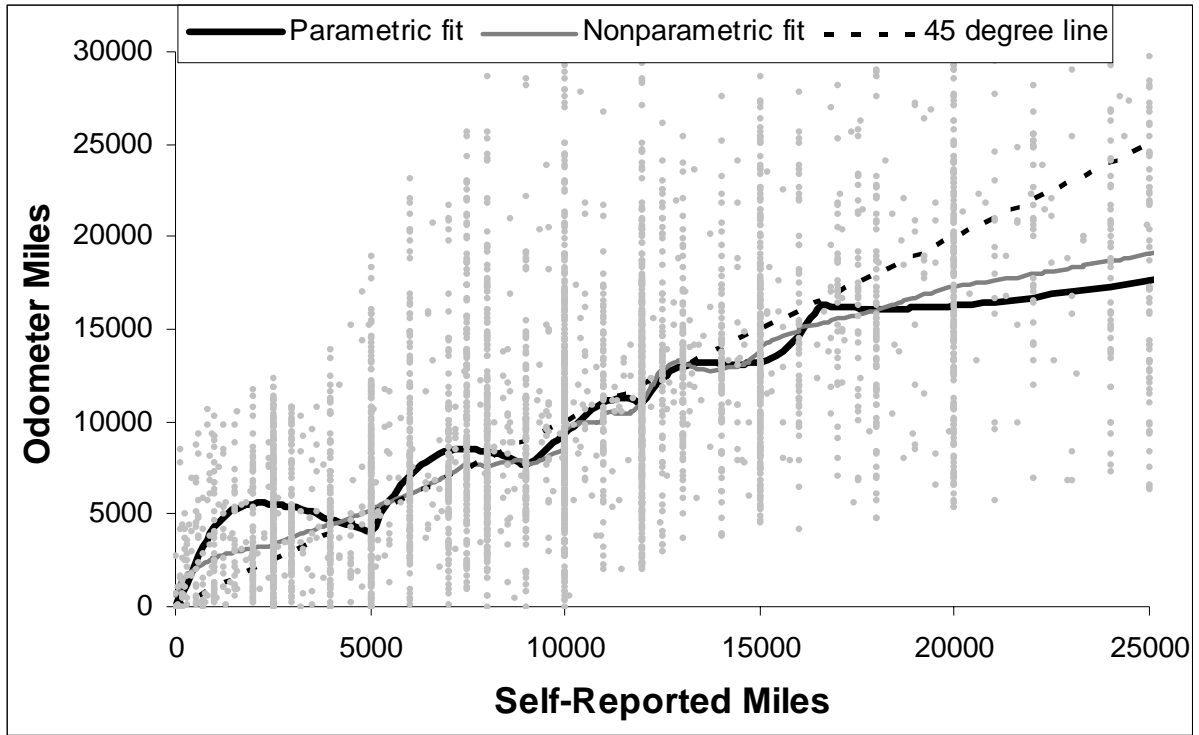


Figure 1: Direct Estimates of Bias in Self-Reported Miles. The parametric fit is from MLE of a piecewise cubic spline with knots at the quartiles of the data. The nonparametric fit is from Friedman’s supersmoother. The scatter plot is the data, which are censored above and below.

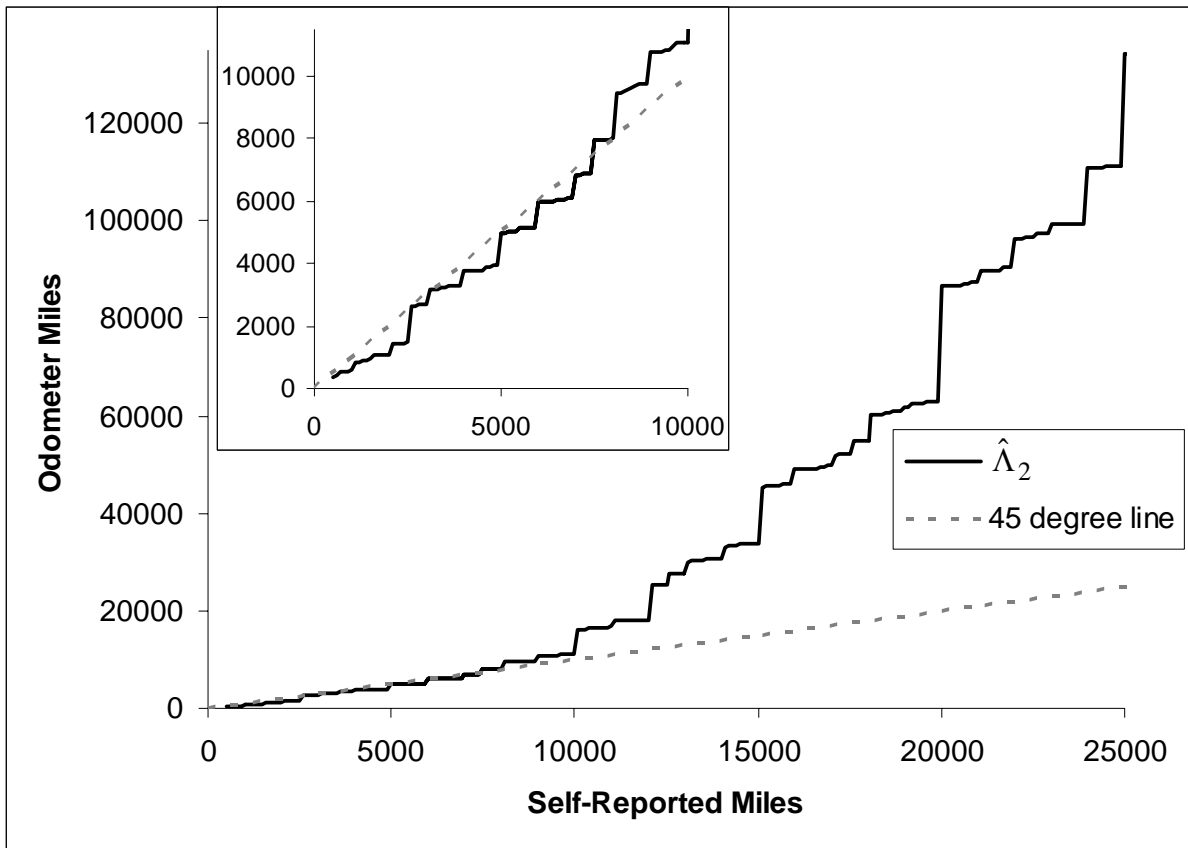


Figure 2: Orbit Estimate of Bias in Self-Reported Miles. Inset graph shows detail for small values of SRMILES.

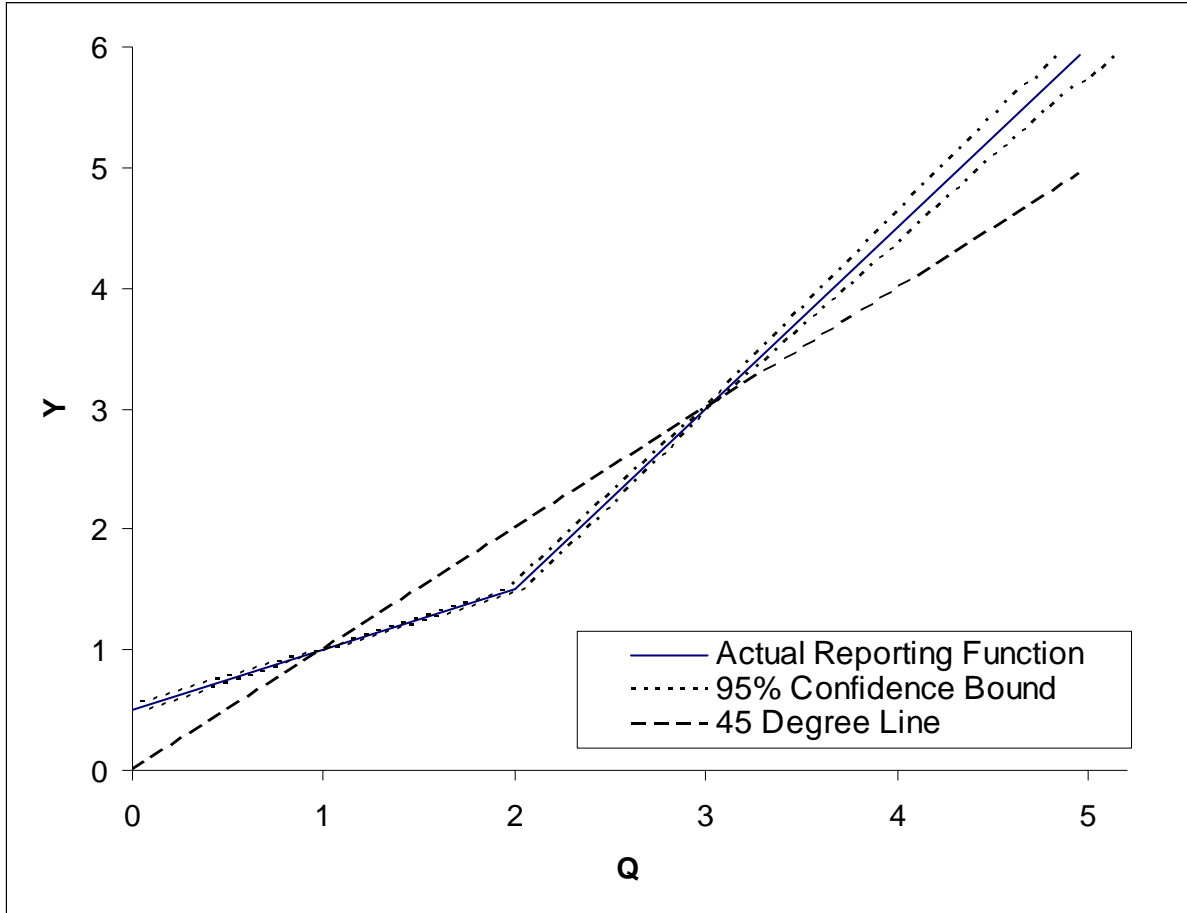


Figure 3: Monte Carlo Results for the Estimate of the Reporting Function. Confidence bounds are calculated pointwise for $\hat{\Lambda}_2(t)$, the inverse of the reporting function. Fixed points are at 1 and 3.

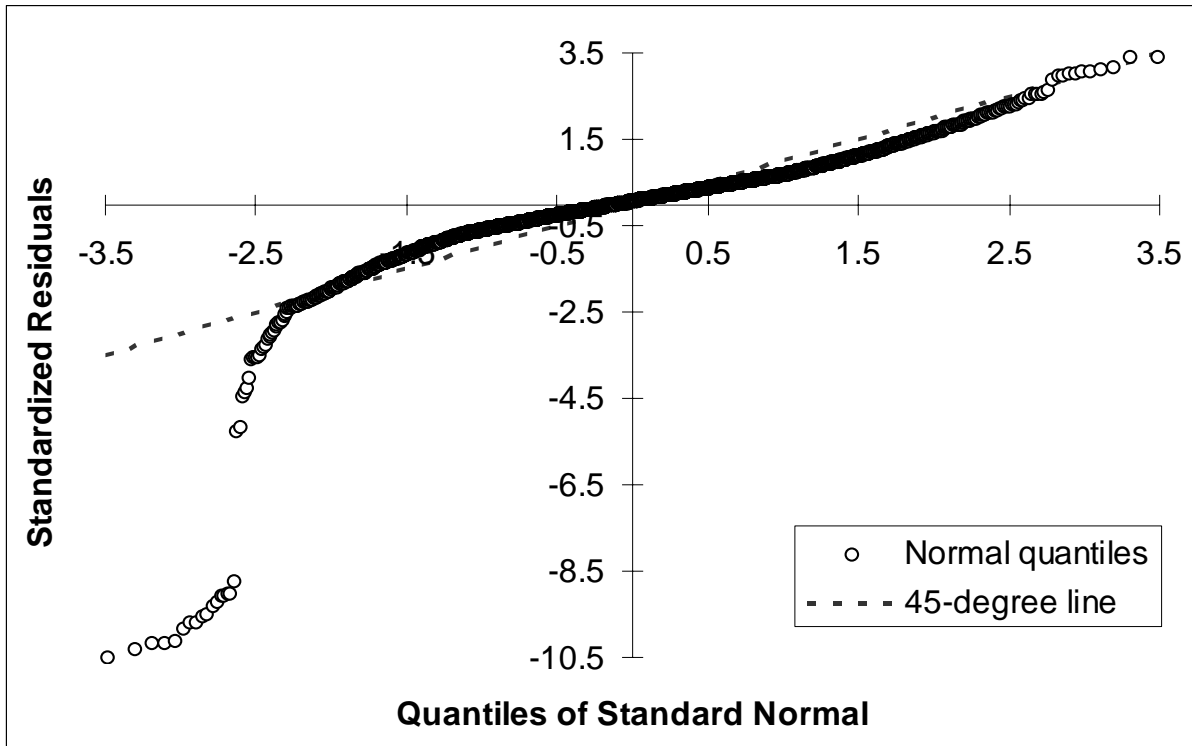


Figure 4: Quantile-Quantile Plot of the Residuals from the ODOMILES Estimation. Points off the 45-degree line indicate departures from normality.

Dep var: Estimator:	ODOMILES		SRMILES		SRMILES	
	MLE		OLS		Orbit	
	<i>coef.</i>	<i>s.e.</i>	<i>coef.</i>	<i>s.e.</i>	<i>coef.</i>	<i>s.e.</i>
Constant	7.67	(0.46)***	7.07	(0.46)***	6.21	(0.90)***
Income1	0.14	(0.04)***	0.23	(0.05)***	0.35	(0.09)***
Income2	-0.08	(0.14)	0.15	(0.10)	0.36	(0.20)*
Income3	0.00	(0.18)	0.17	(0.10)*	0.35	(0.22)
Income4	0.03	(0.22)	0.06	(0.12)	0.26	(0.26)
DriveToWork	0.04	(0.06)	-0.04	(0.04)	0.15	(0.08)*
Carpool	-0.02	(0.14)	-0.10	(0.09)	-0.12	(0.18)
VehicleAge	-0.05	(0.00)***	-0.05	(0.00)***	-0.09	(0.01)***
Van	0.25	(0.08)***	0.15	(0.04)***	0.21	(0.09)**
SUV	0.05	(0.06)	0.07	(0.04)*	0.20	(0.09)**
Pickup	-0.15	(0.06)***	0.01	(0.04)	0.03	(0.09)
Female	-0.13	(0.04)***	-0.27	(0.03)***	-0.50	(0.07)***
Age	0.03	(0.01)***	0.00	(0.01)	0.02	(0.01)**
AgeSq	0.00	(0.00)***	0.00	(0.00)***	0.00	(0.00)***
Employed	0.16	(0.06)**	0.25	(0.05)***	0.15	(0.09)
Rural	0.28	(0.05)***	0.16	(0.03)***	0.28	(0.07)***
Midwest	0.01	(0.06)	0.01	(0.04)	-0.03	(0.08)
South	0.10	(0.05)**	0.09	(0.04)**	0.08	(0.08)
West	0.07	(0.06)	-0.02	(0.04)	-0.02	(0.09)
σ	1.10	(0.00)***	0.85		1.31	(0.06)***
$\ln(L)$	-6074.97		-5259.81		-2314.05	

* = 10% level significance; ** = 5% level significance; *** = 1% level significance.

Table notes: $N = 4,182$. For each pair of columns, the dependent variable is given in the top row and the estimation technique in the second row. The model for *ODOMILES* is given in the appendix. The Orbit estimator for the second *SRMILES* model is presented in (5). *Asy. s.e.* in parentheses. *Incomen* is the income elasticity for the n th household income quartile group.

Table 1: Estimation Results for the Income Elasticity of Annual Mileage

	True Value	Mean Bias	Median Bias	Std. Dev.	95% Confidence Interval
β_0	2.00	3.00E-05	-1.05E-04	2.19E-02	(1.956, 2.042)
β_1	1.00	7.76E-04	-2.20E-05	2.51E-02	(0.953, 1.050)
β_2	-1.00	1.72E-04	7.23E-04	2.54E-02	(-1.051, -0.951)
$\log(\sigma)$	0.00	-1.21E-03	-1.88E-03	2.00E-02	(-0.041, 0.036)

Table notes: $N = 4,000$. Figures are calculated from 999 Monte Carlo runs. 95% Confidence Interval is taken from the empirical density of the Monte Carlo estimates. β_0 is an intercept. Data for covariates corresponding to β_1 and β_2 are generated as described in the text.

Table 2: Monte Carlo Results for Extended Orbit Estimator for θ