

APPENDIX K

CALCULATION OF ANNUALIZED MILEAGE ESTIMATES BASED ON ODOMETER READINGS

Odometer readings for NPTS vehicles were recorded for different time intervals (Table K-1). Mileage differences between odometer readings recorded for individual vehicles reflect driver and household characteristics, as well as seasonal effects on driving.

Table K-1
Time Interval between Two Odometer Readings Recorded for NPTS Vehicles

Percent of NPTS vehicles*	Time interval between two readings
1%	≤ 1½ months
24%	1½ - 2 month
25%	2 - 3¾ months
25%	3¾ - 6 months
20%	6 - 10½ months
5%	10½ - 18⅞ months

* Applied to 42,319 vehicles that have two valid recording dates.

In this appendix, we discuss a method used to "annualize" the number of miles driven between two odometer readings to an estimate of annual driving. In essence, this method adjusts individual vehicle's mileage rates for seasonality. In Section K.1, we discuss data screening necessary before fitting an annualization model and computing annualized estimates. This was an important step, unfortunately, because more than half of the NPTS vehicles were not suitable for this annualization procedure. In Section K.2, the choice of statistical model—a linear model—for the seasonality adjustments is discussed. In Section K.3, we describe the mechanics of computing the annualized estimates as well as standard errors for the estimates. Though brief, part of Section K.3

is technical. Technical background may be found in most any text on linear models, for example, Searle (1971). In Section K.4, we discuss: (1) some adjustments to the annualized driving estimates, and (2) outlier screening and data quality flags based on the annualized estimates. Finally, we outline data-quality limitations in Section K.5.

K.1 Preliminary Data Screening

There were 75,217 vehicles sampled in the 1995 NPTS. Data on many (44%) of them were incomplete, however, in the sense that one or more of the starting and ending odometer readings or one or both of the recording dates were missing. Some of the remaining 56% "complete" observations were anomalous: negative amount of driving between two recording dates, or the difference between odometer readings implying more than 1,440 miles (= 24 hour × 60 miles/hour) of driving per day. About 0.6% of the 75,217 vehicles had a recording period shorter than six weeks, and were excluded from the annualization process since we believe that such short periods would tend to lead to anomalous annualized estimates. Since driver characteristics influence the amount of driving done in the driver's designated vehicle, 5.5 percent of the vehicles were excluded from the annualization calculations because they did not have a designated "primary" driver. Also, motorcycles and vehicles with "other" and "don't know" vehicle types were excluded. As summarized in Table K.2, this screening procedure reduced the original 75,217 vehicles to 36,109 vehicles for which annualized mileage estimates were made.

The NPTS data on odometer mileages and days-of-recording exhibit a lot of variability. This makes annualization difficult, and impacts the quality of the annualized estimates. Among the 36,109 vehicles remaining after the preliminary data screening, 378 (about 1%) had a difference between two odometer readings exceeding 160,000 miles per year and 580 of them had their differences more than 115,000 miles per year. The 115,000 mile figure was considered to be a reasonable upper limit for the annual miles driven in a vehicle, and was used as a cap for the self-reported annual mileage estimates. Users of the annualized estimates should understand the limits imposed by outliers and data variability.

Table K.2 Preliminary Data Screening of the 1995 NPTS Vehicles

Data Problem	Number of Vehicles	Percent
Incomplete data — odometer readings and/or recording dates missing	32,811	43.60
Negative differences between 2 odometer readings	1,040	1.40
Differences between 2 odometer readings too large (more than 1,440 miles per day)	53	0.07
Odometer readings recorded less than six weeks apart	419	0.56
Incomplete data and negative odometer	33	0.04
Negative miles and less than six weeks of data	16	0.02
Mileage too large and less than six weeks of data	5	0.00
No primary driver associated with the vehicle	4,099	5.50
Motorcycles, "other," "don't know" vehicle types	632	0.84
Vehicles with usable data (none of the above)	36,109	48.00
Total 1995 NPTS Vehicles	75,217	100.00

K.2 Choice of Model

The choice of a predictive statistical model should depend on: (1) knowledge of the modeled process; (2) properties of the input data with respect to the number of observations, tendency to have outliers, goodness of model fit, etc.; and (3) mathematical tractability. Mathematical tractability refers to ease of doing computations. Linear models tend to be tractable; nonlinear models can be intractable, for example, because of starting-value or convergence problems. Mathematical tractability is especially important in our application because of the large number of observations and the large number of potential

predictors: education level of the primary driver, MSA size, vehicle age and type, and so on. Because the NPTS data are noisy with respect to the goal of estimating the annual miles of driving based on odometer readings, data variability and the tendency to have outliers are an important consideration. The coefficient of variation of our final prediction model is 1.83, and the (36,109) regression residuals are right skewed, typical of high noise scenarios. While the average of the residuals was of course zero, their 1 and 99 percentiles, for example, were -74.6 and 391.2 miles per year, indicating a wide range of the residuals.

A natural model for the total miles observed for an individual vehicle is

$$\text{total miles} \propto \left(\sum_{\text{day } i} \theta_i \right) \times (\text{factor for class}) \times \text{error}, \quad (1)$$

where "day i " refers to the days in an interval of recording, θ_i is the contribution for day-of-the-year i or perhaps "month-day-of-week" (e.g., January Sunday, November Wednesday); and "factor for class" is a multiplier determined by the class. A class is defined as a particular combination of demographics, vehicle age and type, and other variables. These variables are called *class* variables. The "factor for class" should be greater than one for classes of vehicles in which their primary drivers drive a lot, and less than one for classes of vehicles in which their primary drivers do not drive much. Because a mileage total is modeled here, both the class and error adjustments enter multiplicatively. Because mileages in the NPTS survey were recorded for intervals of varying starting dates and lengths, the summation is needed in (1), rather than a single θ -term, representing an individual month or day. The variable-length intervals thus make annualization more difficult.

Unfortunately, the model (1) is not as tractable as we would like. It is nonlinear. Although appropriate for right-skewed data, a logarithmic transformation does not make the model linear because of the summation. Logarithms may, in any case, be

inappropriate for annualization because they introduce bias. To see this, consider a simple example. Suppose we have just 12 vehicles, each observed for exactly one month, January through December, and suppose there is just one class of vehicles (i.e., these 12 vehicles have identical independent variables). Also suppose there are no day-of-the-week effects, and for simplicity, assume a year is twelve months with exactly thirty days each. Then the annualized mileage per day (mpd) estimate for each vehicle should be the arithmetic mean of the mpd's for all vehicles. On the other hand, if we transform to the log scale, the annualized log mpd estimate for each vehicle would be the arithmetic mean of the log-mpd's for all vehicles. Then the question becomes how we compute the annualized mpd from the annualized log-mpd. If we just take the anti-log of the annualized log-mpd, we get the geometric mean of the mpd's. (The geometric mean is the anti-log of the arithmetic mean of the logs.) It is well-known that the geometric mean is always less than or equal to the arithmetic mean, and that inequality is strict unless all observations are the same. Thus the anti-log of the annualized log-mpd is biased.

If the mpd's were known to be log-normal, we could mathematically correct for the bias. Unfortunately, there is no good basis for assuming log-normality here. In general, there is no way to correct for the bias induced by the log transformation without making some kind of parametric distribution assumption. Thus, although the model (1) is sensible, it has the disadvantage of being nonlinear, not amenable to the log transformation, which would not linearize it anyway, and not very tractable.

To overcome the aforementioned problems, we considered the model

$$\text{rate} = \frac{\text{total miles between 2 readings}}{\text{number of days}} = \text{intercept} + \frac{1}{\text{number of days}} \left(\sum_{\text{month-day } i} \theta_i \right) + (\text{term for class}) + \text{error.} \quad (2)$$

This model **is** linear, and is thus more tractable than model (1). It is similar to (1), but,

because the dependent variable is a rate rather than a total, the additive (rather than multiplicative) adjustments for class and error are reasonable. For the sake of simplicity, we also took θ_i to represent month-day (i.e., month-day-of-week) here rather than day of the year. Thus, for example, if there are two January Sundays in a period of recording, then the θ term for January Sundays would be added in twice. The "number of days" denominator is necessary because the θ 's represent contributions to the total—the more days, the more θ 's—whereas the overall expression is a rate (miles per day).

Here is a simplified example. (A complete example, involving all of the levels of all of the class variables used to fit the model, would be less clear than a simplified one.) Suppose there are just two class variables, say, vehicle age class and vehicle type. Then the class term in our model might be of the form

$$\alpha_i + \beta_j + \gamma_{ij}$$

where α_i is the contribution above the intercept for the i^{th} vehicle age class (main effect of age), β_j is the contribution above the intercept for the j^{th} vehicle type (main effect of vehicle type), and γ_{ij} is the contribution above and beyond the $\alpha_i + \beta_j$ for the i^{th} vehicle age class and the j^{th} vehicle type jointly (two-way interaction of vehicle age and type). Suppose a vehicle's mileage is recorded for January 1-8, 1995 (an overly short interval taken for simplicity). Since 1995 began with a Sunday, this interval represents two January Sundays, and one each for the other January weekdays. If the vehicle age class is "1" (less than one year old) and the vehicle type is "2" (= van), then the model (2) is

$$\text{rate} = \text{intercept} + \frac{1}{8} (2\theta_1 + \theta_2 + \theta_3 + \dots + \theta_7) + \alpha_1 + \beta_2 + \gamma_{12} + \text{error},$$

where $\theta_1, \dots, \theta_7$ are the terms for January days of the week, Sunday through Saturday. Because the model is linear, estimates of the α , β , γ , and θ terms can be computed using software such as the SAS GLM (general linear model) procedure. Then, by revising the

expression involving the θ 's, an annualized rate can be estimated. In the revision, the expression involving the θ 's in the model (2) is changed to

$$\frac{1}{365.25} \sum_{\text{month-day } i} \frac{\text{Days in month of month-day } i}{7} \theta_i,$$

where the sum now extends over all $7 \times 12 = 84$ month-days in a year. These calculations are discussed further in the next subsection.

A model similar to model (2) was derived by Kunert, Hu, and Young (1995) in their analysis of the 1990 NPTS data. Odometer readings were not recorded in the 1990 NPTS. Rather, the amount of driving was recorded for a single designated travel day. Thus, their model had terms to adjust the driving for the particular "travel day." The adjustments in our case are for intervals of, in most cases, many travel days. The class terms in our model were taken from the Kunert et al model, with the following two exceptions: (1) We added terms for the number of drivers in the household. (2) We included all two-way interaction terms. The household driver terms were added on the basis of engineering judgement. Assessing the importance of any of these model terms is difficult. This is because with sample sizes as large as the NPTS data's and with numerous terms for each class variable (because of the interactions) nearly every variable had some statistically significant terms. Fortunately, our primary task here is prediction—annualizing mileage estimates; assessing the importance of the various factors is secondary.

K.3 Computation of the Annualized Estimates

This section contains technical material that may be beyond the interest of the casual reader. The GLM procedure in SAS was used to fit the annualization model. Class variables were education level and age of the primary driver (SAS variable name *educ* and *r_age*, respectively), household composition (*lif_cyc*), vehicle age (created from variable *vehyear*), vehicle type class (*vehtype*), size of MSA (*msasize*), census division (*census_d*),

and household number of drivers per vehicle (created from variables *hhvehcnt* and *drvrcnt*). There are 3,175,000 possible combinations of these classes; obviously not all are represented in the NPTS data. In theory, the two-way interaction model provides some smoothing to adjust out anomalies in low-frequency (i.e., small sample-size) classes.

The multipliers (independent variables) of the terms for "month-day" (the θ -terms) were computed in a preliminary SAS data step. These multipliers were entered into a linear model with all main effects and two-way interactions for the class variables. As an intercept term was included in the model, the last (84th) θ was dropped. (See, for example, Searle, 1971. This reduction to full rank results in no loss of generality; the other independent variables and corresponding parameters are similarly reduced in the GLM algorithm.) The resulting model had 994 degrees of freedom. After data screening (see below), 36,109 observations were used to fit the model, or about 36 observations per degree of freedom (i.e., model parameter).

After fitting the model with SAS' proc GLM, annualized estimates could be computed with it. According to the model,

$$Y = X \hat{\beta} + R,$$

where Y is the vector of observed average daily mileages (based on odometer readings), X is the matrix of independent variables (reduced to full rank), $\hat{\beta}$ is the (reduced) vector of model parameter estimates, and R is the vector of residuals. To "annualize" the observed mileage rates, we simply revise X so that it reflects, for each vehicle, travel for a year rather than for the recording time period for that vehicle. Thus each month-day term is set to

$$\frac{\text{number days in month}}{7 \times 365.25}. \tag{3}$$

With the number of days in February taken to be 28.25, the sum of these terms over days-of-the-week and months (for one year) is 1. Call this modification of X , X^* . With X^* and the same $\hat{\beta}$ (and $X^*\hat{\beta}$ the seasonally adjusted mean) and the residual vector R , a vector of seasonally-adjusted annualized estimates is

$$Y^* = X^* \hat{\beta} + R.$$

To compute the standard errors of these annualized estimates, notice that

$$\begin{aligned} Y^* &= X^* \hat{\beta} + R = X^*(X'X)^{-1}X'Y + (I-P)e \\ &= X^*\beta + X^*(X'X)^{-1}X'e + (I-P)e = X^*\beta + (P^* + I - P)e, \end{aligned}$$

where $P = X(X'X)^{-1}X'$, $P^* = X^*(X'X)^{-1}X'$, β is the "true" parameter vector, and e is the vector of errors ($Y = X\beta + e$). Here "' " denotes matrix transpose. We have also used here the fact that $R = (I - P)e$. Therefore (using a property of the variance of linear functions), where V denotes the variance of an individual y -value (daily mileage rate),

$$\text{Cov}(Y^*) = V(P^* + I - P)(P^* + I - P)' = V(P^*P^{*'} + P^*(I - P) + (I - P)P^{*'} + I - P).$$

It is straightforward to verify that $P^*(I - P) = 0$. It follows that

$$\text{Cov}(Y^*) = VP^*P^{*'} + V(I - P),$$

and that for y^* an element of Y^* and x^* and r , the corresponding elements of X^* and R ,

$$\text{stderr}(y^*) = [(\text{stderr}(x^*\hat{\beta}))^2 + (\text{stderr}(r))^2]^{1/2}.$$

The standard error of y^* is straightforward to compute in SAS, because $stderr(x^*\hat{\beta})$ is the standard error of a predicted mean value, and $stderr(r)$ is the standard error of a residual, both of which can be output directly with proc GLM.

The above seasonally-adjusted daily mileage rates and their standard errors were converted to annual rates (miles driven per year) and standard errors by multiplying them by 365.25. In addition to these annualized estimates (SAS variable *annualzd*) and standard errors (*stderr*), alternative "crude" estimates (*mtd365*) were computed by multiplying 365.25 by each crude daily rate (i.e., the difference between odometer readings for a vehicle divided by the number of days in the recording period of that vehicle.) Standard errors (*std365*) for these estimates were also computed, as above, except no month-day terms were included in the linear model. Crude mileage estimates and standard errors can likewise be computed for any time period, in particular, the periods for which the odometer readings were taken.

K.4 Outlier Screening

Despite the extensive preliminary data screening, the remaining data and annualized estimates are noisy. Certain common-sense restrictions are violated. For example, some of the annualized estimates are less than the difference between odometer readings (for periods of less than one year). Some of the annualized estimates are negative. To understand how this can happen, remember that the dependent variable of the model is a daily **rate** (odometer mileage per day of recording). The annualized daily rate can easily be less than the crude daily rate of the dependent variable, and, especially when the corresponding residual is negative and large, the annualized rate can be less than the difference between two odometer readings itself. The model has no constraint to automatically prevent this.

Estimates that violated common-sense restrictions were adjusted as follows. For vehicles whose recording period was less than one year, if the annualized estimate was

less than the difference between two odometer mileage (this includes negative estimates), the annualized estimate was set to be the difference between two odometer readings itself. For any annualized estimate whose recording period was more than 365 days, a negative annualized estimate was set to the crude estimate (*mtd365*), and an annualized estimate greater than the corresponding difference between two odometer readings was set to be the difference between two odometer readings. Also, annualized estimates greater than 115,000 were set to be 115,000. This cap was set to be consistent with the cap used on the self-reported estimates of annual driving (*annmiles*). These changes were made with the following frequencies.

Table K-3. Codes for Adjustments to Annualized Estimates of Driving

Code	Frequenc	Percent	Meaning
(no code)	32,289	89.4	No adjustment was made
1	3,800	10.5	Number of days between two readings less than 366, and annualized estimate less than difference between odometer readings; annualized set to difference between odometer readings.
2	16	.0	Number of days between two readings greater than 365, and annualized estimate greater than difference between odometer readings; annualized set to difference between odometer readings.
3	4	.0	Number of days between two readings greater than 365, and annualized estimate negative; annualized set to crude estimate*.
Total	36,109	100.0	(All)

*The crude estimate is 365.25 times the odometer difference divided by days in observation period.

Although adjustments of Code 1 had to be made for 3,800 household vehicles, the adjustments were minor in nearly all cases, amounting to less than 2,000 miles for all but 799 household vehicles, and less than 5,000 miles for all but 111 vehicles (.3% of 36,109). (A SAS variable *ann_edit* flags these adjustments, though per a modification discussed in the next section.)

After making these adjustments, each adjusted annualized estimate was compared to its "crude" analog (*mtd365*) and to a corresponding self-reported estimate (annual miles driven reckoned by driver). Outlier codes were then assigned on the basis of these comparisons and subjectively determined thresholds (Table K-4). Because the self-reported estimates were considered less reliable than the crude estimates, the thresholds are tighter for the crude-vs-annualized comparisons. Codes based on comparisons of the

annualized and the crude estimates were only assigned if the difference exceeded 5,000 miles. Codes based on comparisons of the annualized and the self-reported estimates were only assigned if the difference exceeded 10,000 miles. The outlier codes are recorded as numeric codes (SAS variable *ann_out*) as indicated in Table K-4. Out of the 36,109 vehicles whose annual miles driven were estimated based on their odometer readings, 32,153 (89%) are considered to have reasonable annualized estimates (i.e., not outliers).

Table K-4. Outliers Codes of Annualized Estimates of Driving

Code	Numeric Code (for SAS output)	Frequency	Percent	Criteria
(no code)	(no code)	32,153	89.0	Not an outlier
a	2	1,164	3.2	Annualized ^a < Reported ^b / 4 and Annualized - Reported > 10,000
b	5	2,293	6.3	Annualized > 4 × Reported and Annualized - Reported > 10,000
A	1	336	0.9	Annualized < Crude ^c / 2 and Annualized - Crude > 5,000
Aa	3	83	0.2	(A and a)
B	4	75	0.2	Annualized > 2 × Crude and Annualized - Crude > 5,000
Bb	6	5	0.0	(B and b)
Total		36,109	100.0	(all)

^a Estimates of annual driving based on two odometer readings (*annualzd*).

^b Driver self-reported annual mileage estimate (*annmiles*).

^c 365.25 times the difference between odometer readings divided by days in observation time interval (*mtd365*).

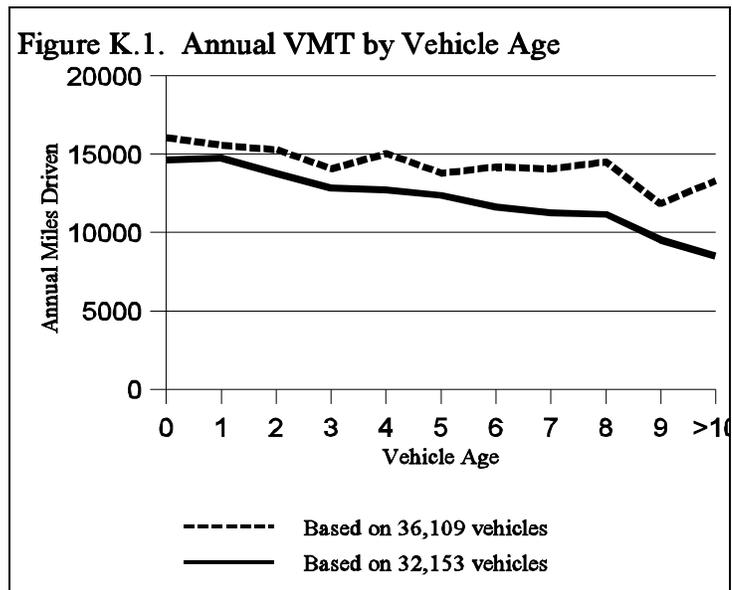
K-5 Limitations

The outlier flags in Table K-4 could indicate either data quality problems or issues pertinent to the annualization model. Data quality problems are those embedded in the information collected from the survey respondents. Issues pertinent to the model are those resulting from the annualization process. As previously mentioned, there **are** data quality problems. Most of the time the flags indicate such problems. To illustrate more generally the magnitude of these data problems, we calculated the correlation between the annualized and crude estimates to be 0.998. Thus there is very good agreement between the annualized estimates and the actual data (i.e., differences between two odometer readings). However, the correlation between either the annualized or crude estimates and the self-reported estimates is only 0.11, indicating that the self-reported miles driven in a year bear little relationship to the annual miles driven estimated based on the odometer readings. Now, if we restrict attention to the 32,153 observations that were not assigned any of the outlier flags in Table K-4, then the correlation between the annualized and the self-reported estimates increases considerably to 0.62. This implies that if we remove the problematic data, then the self-reported miles driven in a year relate significantly more to the annual miles driven estimated based on the odometer readings than if problematic data were included in the calculation

(0.62 vs. 0.11, respectively). This illustrates that the magnitude of the data quality problems is substantial compared to the issues related to the annualization process.

For another example of data quality problems, we compare the average annual miles driven per vehicle (i.e., VMT) by age of the vehicle (Figure K.1). The first set of averages are for all 36,109

annualized estimates with a mileage cap of 115,000, while the second set are for the



32,153 unflagged annualized estimates.

For the 32,153 unflagged estimates, the steadily decreasing trend of annual miles driven with vehicle age seems much more consistent with those observed in other data sources than the corresponding, much less even, results for the 36,109 vehicles. In these data, the cap was used to deal with anomalous, high mileages. Without the mileage-cap, the comparison becomes even more polar. For this reason, annualized estimates that exceeded 115,000 miles were capped at 115,000 in the final NPTS data set. Quality flags (*ann_edit*) in the final NPTS data set are summarized in Table K-5. To maintain reasonable analysis results, users are urged not to overlook these data quality flags.

Table K-5 Final Codes for Adjustments to the Final Annualized Estimates

Code	Frequenc	Percent	Criteria
(no code)	31,721	87.8	No adjustment
1	3,799	10.5	Number of days less than 366, and annualized estimate less than difference between odometer readings; annualized set to odometer difference.
2	16	.0	Number of days greater than 365, and annualized estimate greater than difference between odometer readings; annualized set to odometer difference.
3	4	.0	Number of days greater than 365, and annualized estimate negative; annualized set to crude estimate*.
4	568	1.6	None of above, but mileage exceeds 115,000; capped at 115,000 miles.
5	1	.0	As in 1 above, and capped at 115,000
Total	36,109	100.0	(All)

* The crude estimate is 365.25 times the odometer difference divided by the number of days in the reporting period.

References: Kunert, U., Hu, P., and Young, J. (1995). "Framework for the Expansion and the Analysis of the 1995 Nationwide Personal Transportation Survey Odometer Reading Data," (unpublished report).

Searle, S. R. (1971). *Linear Models*, John Wiley & Sons, New York.