Modelling Annual VKT Considering Heterogeneous Approximation Biases

Toshiyuki Yamamoto
Noritaka Nakagawa
Outline

- Background
- Objective
- Data
- Analysis
- Future tasks
Annual vehicle kilometres travelled

VKT (vehicle kilometres travelled)
- has been used as an index of car use
  - The strongest indicator of car dependencies and household’s travel patterns
- There have been many studies to make use of VKT for various purposes
  - Gasoline consumption, vehicle emissions, and crashes
Difficulty in modelling VKT

Generally, goodness-of-fit is low

- $R^2$: 0.11 (Train, 1986), 0.15 (Kockelman, 1997), 0.17 (Yamamoto et al., 2001)

Reason might be

- Variability among household’s vehicle use
  - Factors to affect car use are not fully incorporated
- Inaccuracy in observation
  - Annual VKT reported by respondents
  - Short-period odometer readings
Outline

- Background
- Objective
- Data
- Analysis
- Future tasks
Objectives

- Inaccuracy in observation is examined
- Annual VKT model is developed considering inaccuracy in observation
  - Efficiency is compared with conventional models
- Heterogeneity among respondents in inaccuracy of observation is also examined
Incomplete data

- **Missing data**: each data value is either perfectly known or entirely unknown
- **Coarse data**: only a subset of the complete-data sample space is observed
  - **Rounding**: data value is observed only to the nearest integer
  - **Censoring**: in failure time data, if an item has not failed by the time observation ends, failure time is known only to lie beyond the last observation point
**Heaping**

- one of the coarsening related with rounding.
- includes the phenomenon known as digit preference.
- includes items reported with various levels of coarseness
- E.g., histograms of age often exhibit heaps at common ages such as integral multiples of ten years with adults, or integral multiples of six or twelve months with children.
Coarseness in VKT data

- Annual VKT reported by respondents includes some level of approximation
- Level of approximation may vary among respondents

VKT data is regarded as heaped
Outline

- Background
- Objective
- Data
- Analysis
- Future tasks
Parc-Auto

- French households’ car ownership panel data
- Conducted yearly since 1976, and continues today
- Sample size is maintained at about 7,000 households each year
- Includes characteristics of up to 3 cars in the household, vehicle use, general attitudes concerning transportation, etc.
VKT data in Parc-Auto

2 types of information
- Difference in odometer readings at 2 successive years -> Calculated VKT
- Annual mileage in kilometres reported by respondent -> Reported VKT

We use for analysis 1167 sample cases
- 1998 VKT data
- Sub-sample who answered both 1997 & 1998 survey to get Calculated VKT
Sample distribution

- Reported VKT is obviously heaped

Calculated VKT

Reported VKT
Relationship between calculated and reported VKTs

Some cases have large discrepancies
Relationship between calculated and reported VKTs

\[ y = 0.9143x + 814.7 \]

\[ R^2 = 0.8151 \]

Cases with 10000+ difference are discarded.
Outline

- Background
- Objective
- Data
- Analysis
- Future tasks
Procedure of analysis

1. Regression models of both calculated and reported VKT
2. Regression models of error in reporting assuming that calculated VKT is true
3. Ordered-probit models of VKT using pre-determined coarseness
4. Ordered-probit model of possible maximum coarseness of the report by each respondent (heterogeneity among respondent)
5. Latent class model of VKT considering heterogeneity in coarseness among respondents
Explanatory variables

- Household’s attribute
  - #children (15-), PT access., large city (300,000+), #cars, low income (F75,000-), high income (F200,000+)

- Personal attribute
  - Young (39-), old (60+), worker, male, car commute

- Car attribute
  - Diesel car, small car, large car, truck, car age
1. Regression models of VKT

- Unexpectedly, R2 is higher in reported VKT model
- Most var. have similar significance in both models

<table>
<thead>
<tr>
<th></th>
<th>Calculated</th>
<th></th>
<th></th>
<th>Reported</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef.</td>
<td>t-stat.</td>
<td>Coef.</td>
<td>t-stat.</td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>9.380</td>
<td>70.9</td>
<td>9.345</td>
<td>81.5</td>
<td></td>
</tr>
<tr>
<td>#children</td>
<td>-0.015</td>
<td>-0.5</td>
<td>-0.014</td>
<td>-0.5</td>
<td></td>
</tr>
<tr>
<td>PT access</td>
<td>-0.070</td>
<td>-1.4</td>
<td>-0.061</td>
<td>-1.4</td>
<td></td>
</tr>
<tr>
<td>Large city</td>
<td>0.116</td>
<td>2.1</td>
<td>0.013</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>#cars</td>
<td>-0.038</td>
<td>-0.9</td>
<td>0.017</td>
<td>0.5</td>
<td></td>
</tr>
<tr>
<td>Low income</td>
<td>-0.115</td>
<td>-1.4</td>
<td>-0.202</td>
<td>-2.8</td>
<td></td>
</tr>
<tr>
<td>High income</td>
<td>0.041</td>
<td>0.8</td>
<td>0.057</td>
<td>1.3</td>
<td></td>
</tr>
<tr>
<td>Young</td>
<td>0.095</td>
<td>1.4</td>
<td>0.105</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>Old</td>
<td>-0.326</td>
<td>-3.8</td>
<td>-0.201</td>
<td>-2.7</td>
<td></td>
</tr>
<tr>
<td>Worker</td>
<td>-0.137</td>
<td>-1.7</td>
<td>-0.074</td>
<td>-1.0</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>0.115</td>
<td>2.3</td>
<td>0.110</td>
<td>2.6</td>
<td></td>
</tr>
<tr>
<td>Car commute</td>
<td>0.390</td>
<td>6.3</td>
<td>0.357</td>
<td>6.6</td>
<td></td>
</tr>
<tr>
<td>Diesel car</td>
<td>0.389</td>
<td>8.2</td>
<td>0.379</td>
<td>9.2</td>
<td></td>
</tr>
<tr>
<td>Small car</td>
<td>-0.269</td>
<td>-5.3</td>
<td>-0.174</td>
<td>-3.9</td>
<td></td>
</tr>
<tr>
<td>Large car</td>
<td>0.163</td>
<td>2.1</td>
<td>0.160</td>
<td>2.4</td>
<td></td>
</tr>
<tr>
<td>Truck</td>
<td>0.536</td>
<td>1.5</td>
<td>0.579</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>Car age</td>
<td>-0.037</td>
<td>-6.6</td>
<td>-0.038</td>
<td>-7.9</td>
<td></td>
</tr>
<tr>
<td>s.e.</td>
<td>0.634</td>
<td></td>
<td>0.550</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td></td>
<td>0.324</td>
<td></td>
<td>0.347</td>
<td></td>
</tr>
<tr>
<td>Sample size</td>
<td></td>
<td>975</td>
<td></td>
<td>975</td>
<td></td>
</tr>
</tbody>
</table>
Distribution of error in reporting

- (Calculated VKT) – (Reported VKT)
Distribution of **absolute** error in reporting
2. Regression models of error in reporting

- R2 are low in both models
- Calculated VKT has a highly significant coef.

<table>
<thead>
<tr>
<th></th>
<th>W/O calculated VKT</th>
<th>W calculated VKT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef.</td>
<td>t-stat.</td>
</tr>
<tr>
<td>Intercept</td>
<td>766</td>
<td>1.119</td>
</tr>
<tr>
<td>#children</td>
<td>4</td>
<td>0.0</td>
</tr>
<tr>
<td>PT access</td>
<td>96</td>
<td>0.4</td>
</tr>
<tr>
<td>Large city</td>
<td>499</td>
<td>1.7</td>
</tr>
<tr>
<td>#cars</td>
<td>-140</td>
<td>-0.6</td>
</tr>
<tr>
<td>Low income</td>
<td>561</td>
<td>1.3</td>
</tr>
<tr>
<td>High income</td>
<td>-173</td>
<td>-0.7</td>
</tr>
<tr>
<td>Young</td>
<td>-75</td>
<td>-0.2</td>
</tr>
<tr>
<td>Old</td>
<td>-1396</td>
<td>-3.2</td>
</tr>
<tr>
<td>Worker</td>
<td>-1054</td>
<td>-2.5</td>
</tr>
<tr>
<td>Male</td>
<td>-210</td>
<td>-0.8</td>
</tr>
<tr>
<td>Car commute</td>
<td>136</td>
<td>0.4</td>
</tr>
<tr>
<td>Diesel car</td>
<td>-78</td>
<td>-0.3</td>
</tr>
<tr>
<td>Small car</td>
<td>-508</td>
<td>-1.9</td>
</tr>
<tr>
<td>Large car</td>
<td>-779</td>
<td>-2.0</td>
</tr>
<tr>
<td>Truck</td>
<td>-1184</td>
<td>-0.6</td>
</tr>
<tr>
<td>Car age</td>
<td>33</td>
<td>1.2</td>
</tr>
<tr>
<td>Calculated VKT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s.e.</td>
<td>3250</td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>0.016</td>
<td></td>
</tr>
<tr>
<td>Sample size</td>
<td>965</td>
<td></td>
</tr>
</tbody>
</table>
2. Regression models of absolute error

- Logarithm of absolute error is used as dependent variable
- R2 are low in both models
- No significant var. except #cars

<table>
<thead>
<tr>
<th>Var</th>
<th>W/O calculated VKT</th>
<th>W calculated VKT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef.</td>
<td>t-stat.</td>
</tr>
<tr>
<td>Intercept</td>
<td>7.513</td>
<td>25.4</td>
</tr>
<tr>
<td>#children</td>
<td>0.033</td>
<td>0.5</td>
</tr>
<tr>
<td>PT access</td>
<td>0.126</td>
<td>1.1</td>
</tr>
<tr>
<td>Large city</td>
<td>-0.007</td>
<td>-0.1</td>
</tr>
<tr>
<td>#cars</td>
<td>-0.225</td>
<td>-2.4</td>
</tr>
<tr>
<td>Low income</td>
<td>-0.037</td>
<td>-0.2</td>
</tr>
<tr>
<td>High income</td>
<td>0.002</td>
<td>0.0</td>
</tr>
<tr>
<td>Young</td>
<td>-0.044</td>
<td>-0.3</td>
</tr>
<tr>
<td>Old</td>
<td>-0.142</td>
<td>-0.7</td>
</tr>
<tr>
<td>Worker</td>
<td>-0.020</td>
<td>-0.1</td>
</tr>
<tr>
<td>Male</td>
<td>-0.079</td>
<td>-0.7</td>
</tr>
<tr>
<td>Car commute</td>
<td>0.163</td>
<td>1.2</td>
</tr>
<tr>
<td>Diesel car</td>
<td>0.192</td>
<td>1.8</td>
</tr>
<tr>
<td>Small car</td>
<td>-0.204</td>
<td>-1.8</td>
</tr>
<tr>
<td>Large car</td>
<td>0.105</td>
<td>0.6</td>
</tr>
<tr>
<td>Truck</td>
<td>0.129</td>
<td>0.2</td>
</tr>
<tr>
<td>Car age</td>
<td>-0.018</td>
<td>-1.4</td>
</tr>
<tr>
<td>Calculated VKT</td>
<td>0.103</td>
<td>1.4</td>
</tr>
<tr>
<td>s.e.</td>
<td>1.406</td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>0.010</td>
<td></td>
</tr>
<tr>
<td>Sample size</td>
<td>965</td>
<td></td>
</tr>
</tbody>
</table>
3. Ordered-probit models of VKT using pre-determined coarseness

- Reports are assumed as rounded as multiples of 500, 1000, or 5000 km regardless of reported value.

\[
\ln(VKT^*) = \beta X + \varepsilon
\]

\(VKT\) is observed if \(VKT - \delta \leq VKT^* < VKT + \delta\)

\[
\Pr(VKT) = \Phi\left(\frac{\ln(VKT + \delta) - \beta X}{\sigma}\right) - \Phi\left(\frac{\ln(VKT - \delta) - \beta X}{\sigma}\right)
\]

\(\delta): 500, 1000, or 5000

\(\sigma): \text{standard error (estimated while thresholds are fixed in this model)}\)
1000 km rounding

Compared with regression model,

- explanatory var. have similar coef. estimates
- s.e. has a smaller value

<table>
<thead>
<tr>
<th></th>
<th>Regression</th>
<th>Coef.</th>
<th>t-stat.</th>
<th>1000 km rounded</th>
<th>Coef.</th>
<th>t-stat.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>9.345</td>
<td>81.5</td>
<td></td>
<td>9.340</td>
<td>88.8</td>
<td></td>
</tr>
<tr>
<td>#children</td>
<td>-0.014</td>
<td>-0.5</td>
<td></td>
<td>-0.016</td>
<td>-0.6</td>
<td></td>
</tr>
<tr>
<td>PT access</td>
<td>-0.061</td>
<td>-1.4</td>
<td></td>
<td>-0.053</td>
<td>-1.3</td>
<td></td>
</tr>
<tr>
<td>Large city</td>
<td>0.013</td>
<td>0.3</td>
<td></td>
<td>0.026</td>
<td>0.6</td>
<td></td>
</tr>
<tr>
<td>#cars</td>
<td>0.017</td>
<td>0.5</td>
<td></td>
<td>0.011</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>Low income</td>
<td>-0.202</td>
<td>-2.8</td>
<td></td>
<td>-0.197</td>
<td>-2.9</td>
<td></td>
</tr>
<tr>
<td>High income</td>
<td>0.057</td>
<td>1.3</td>
<td></td>
<td>0.057</td>
<td>1.4</td>
<td></td>
</tr>
<tr>
<td>Young</td>
<td>0.105</td>
<td>1.8</td>
<td></td>
<td>0.103</td>
<td>1.9</td>
<td></td>
</tr>
<tr>
<td>Old</td>
<td>-0.201</td>
<td>-2.7</td>
<td></td>
<td>-0.187</td>
<td>-2.7</td>
<td></td>
</tr>
<tr>
<td>Worker</td>
<td>-0.074</td>
<td>-1.0</td>
<td></td>
<td>-0.067</td>
<td>-1.0</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>0.110</td>
<td>2.6</td>
<td></td>
<td>0.100</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>Car commute</td>
<td>0.357</td>
<td>6.6</td>
<td></td>
<td>0.344</td>
<td>6.9</td>
<td></td>
</tr>
<tr>
<td>Diesel car</td>
<td>0.379</td>
<td>9.2</td>
<td></td>
<td>0.373</td>
<td>9.8</td>
<td></td>
</tr>
<tr>
<td>Small car</td>
<td>-0.174</td>
<td>-3.9</td>
<td></td>
<td>-0.173</td>
<td>-4.2</td>
<td></td>
</tr>
<tr>
<td>Large car</td>
<td>0.160</td>
<td>2.4</td>
<td></td>
<td>0.159</td>
<td>2.5</td>
<td></td>
</tr>
<tr>
<td>Truck</td>
<td>0.579</td>
<td>1.8</td>
<td></td>
<td>0.564</td>
<td>1.9</td>
<td></td>
</tr>
<tr>
<td>Car age</td>
<td>-0.038</td>
<td>-7.9</td>
<td></td>
<td>-0.036</td>
<td>-8.0</td>
<td></td>
</tr>
<tr>
<td>s.e.</td>
<td>0.550</td>
<td></td>
<td>43.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log-likelihood</td>
<td>-3048.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample size</td>
<td>975</td>
<td></td>
<td>975</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Comparison of goodness-of-fit

Standard error (Log-likelihood* in parenthesis)

<table>
<thead>
<tr>
<th></th>
<th>Regression</th>
<th>500 km</th>
<th>1000 km</th>
<th>5000 km</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculated VKT</td>
<td>0.634</td>
<td>0.603</td>
<td>0.583</td>
<td>0.526</td>
</tr>
<tr>
<td></td>
<td>(0.324)*</td>
<td>(-3826)</td>
<td>(-3133)</td>
<td>(-1565)</td>
</tr>
<tr>
<td>Reported VKT</td>
<td>0.550</td>
<td>0.533</td>
<td>0.508</td>
<td>0.475</td>
</tr>
<tr>
<td></td>
<td>(0.347)*</td>
<td>(-3753)</td>
<td>(-3049)</td>
<td>(-1492)</td>
</tr>
</tbody>
</table>

* R2 is shown for regression model

- Smaller standard errors are estimated when larger rounding is assumed
Sample distribution of possible maximum coarseness

500 km rounding: 5500, 10500, …
1000 km rounding: 6000, 11000, …
5000 km rounding: 5000, 10000, …
Sample distribution of possible maximum coarseness

<table>
<thead>
<tr>
<th></th>
<th>Pooled</th>
<th>500 km</th>
<th>1000 km</th>
<th>5000 km</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>1919</td>
<td>163</td>
<td>786</td>
<td>732</td>
</tr>
<tr>
<td>Male</td>
<td>69.4%</td>
<td>71.8%</td>
<td>69.8%</td>
<td>68.7%</td>
</tr>
<tr>
<td>Young</td>
<td>19.5%</td>
<td>13.5%</td>
<td>18.1%</td>
<td>25.5%</td>
</tr>
<tr>
<td>Old</td>
<td>47.7%</td>
<td>62.6%</td>
<td>48.7%</td>
<td>37.8%</td>
</tr>
<tr>
<td>Low income</td>
<td>7.2%</td>
<td>7.4%</td>
<td>5.9%</td>
<td>8.2%</td>
</tr>
<tr>
<td>High income</td>
<td>25.7%</td>
<td>23.9%</td>
<td>25.4%</td>
<td>27.0%</td>
</tr>
<tr>
<td>Large city</td>
<td>82.6%</td>
<td>84.7%</td>
<td>81.2%</td>
<td>84.0%</td>
</tr>
<tr>
<td>Small car</td>
<td>28.8%</td>
<td>40.5%</td>
<td>28.9%</td>
<td>24.5%</td>
</tr>
<tr>
<td>Large car</td>
<td>8.5%</td>
<td>4.9%</td>
<td>8.9%</td>
<td>9.4%</td>
</tr>
</tbody>
</table>
Sample distribution of coarseness admitting multiple possibilities

<table>
<thead>
<tr>
<th></th>
<th>Pooled</th>
<th>500 km</th>
<th>1000 km</th>
<th>5000 km</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>1919</td>
<td>1681</td>
<td>1518</td>
<td>732</td>
</tr>
<tr>
<td>Male</td>
<td>69.4%</td>
<td>69.5%</td>
<td>69.3%</td>
<td>68.7%</td>
</tr>
<tr>
<td>Young</td>
<td>19.5%</td>
<td>20.9%</td>
<td>21.7%</td>
<td>25.5%</td>
</tr>
<tr>
<td>Old</td>
<td>47.7%</td>
<td>45.3%</td>
<td>43.5%</td>
<td>37.8%</td>
</tr>
<tr>
<td>Low income</td>
<td>7.2%</td>
<td>7.0%</td>
<td>7.0%</td>
<td>8.2%</td>
</tr>
<tr>
<td>High income</td>
<td>25.7%</td>
<td>26.0%</td>
<td>26.2%</td>
<td>27.0%</td>
</tr>
<tr>
<td>Large city</td>
<td>82.6%</td>
<td>82.7%</td>
<td>82.5%</td>
<td>84.0%</td>
</tr>
<tr>
<td>Small car</td>
<td>28.8%</td>
<td>28.1%</td>
<td>26.7%</td>
<td>24.5%</td>
</tr>
<tr>
<td>Large car</td>
<td>8.5%</td>
<td>8.7%</td>
<td>9.2%</td>
<td>9.4%</td>
</tr>
</tbody>
</table>
Outline

- Background
- Objective
- Data
- Analysis
- Future tasks
Ordered-probit model of possible maximum coarseness of the report

- Conventional ordered-probit model with categories (500-, 1000, 5000+) will be estimated

- Sample distribution of possible maximum coarseness suggest that age, income, and car size may have significant effects
Latent class model of VKT considering heterogeneity in coarseness among respondents

- Latent class represents the level of coarseness in the report
  - Reported VKT of 6000 km belongs to both 500 km and 1000 km rounding classes
  - 12500 km only belongs to 500 km rounding class

- Ordered-probit model is applied to represent the class

- VKT is also developed as ordered-probit model