観測の不完全性を考慮した自動車年間走行距離の分析*

Analysis on Annual Vehicle Kilometers Traveled Considering Observation Errors*

山本俊行**・中川展孝***

By Toshiyuki YAMAMOTO**· Noritaka NAKAGAWA***

1. はじめに

自動車の走行距離は，日々の交通行動の一環である自動車利用を一定期間内で集計したものであり，その一連の行動を簡略にモデル化したものが走行距離モデルである．これまで自動車の走行距離モデルは，1年間や1ヶ月単位での走行距離を被説明変数とし，世帯の自動車利用状況の記述のために用いられてきた．しかし多くの既存研究では，年間走行距離の予測精度は低いレベルに留まっている．その理由の一つには，分析者がモデルの特定化ならびに分析に際しての確率分布の仮定等を適切に行えていないことが考えられる．もう一つの理由として，観測されたデータが不完全である可能性が考えられる．

年間走行距離の観測は，主に，被験者の記憶に頼った回顧データ，あるいは，1週間等の一定期間の交通行動履歴，オドメーターの読み取り，によって観測される．回顧データを用いる場合には，被験者が正確に走行距離を把握しているとは考えにくく，被験者の回答には何らかの誤差が含まれていると考えられる.一方,1週間等の期間を対象とした場合には,年間走行距離の算出にあたって得られた数値を拡大する必要がある．その際，データとして収集した1週間が平均的な1週間とは限らず,季節変動も大きいと考えられるため，正確な拡大が困難である．

本研究では，より一般的に用いられている，回顧データを対象として，年間走行距離の報告における報告誤差に関する分析を行う．また，報告誤差の存在を考慮することによって，より予測精度の高い年間走行距離モデルの構築を目指す．

*キーワーズ：自動車保有・利用，不完全データ
**正員，博(工)，名古屋大学大学院工学研究科
（愛知県名古屋市千種区不老町，TEL:052-789-4636,
E-mail:yamamoto@civil.nagoya-u.ac.jp）
***正員，修(工)，JR四国（香川県高松市浜ノ町8-33,
TEL: 087-825-1622）

2. データ

本研究で用いるデータは1976年から今日に至るまでフランスで毎年実施されている「Parc Auto」と呼ばれるパネル調査データである．毎年，約7000世帯のサンプル数を維持しており,4年毎にサンプルが入れ替わっていくローテーションパネルの形態を取っている．世帯属性，個人属性，保有車種属性，自動車利用状況等の詳細な情報を含んでおり利用価値の高いデータベースとなっている．

本研究では，このうち，1997年と1998年の両方に回答した世帯のデータを用いる．年間走行距離も調査項目に含まれているため，被験者が調査時点で保有している車両の過去1年間の走行距離を回答した数値が得られている(以降では回答値と呼ぶ).一方,データには,調査時点の各保有車両の総走行距離(オドメーター値）も含まれているため，連続する2回の調査で同じ保有自動車のオドメーター値の差から年間走行距離を計算することが出来る（以降では計算値と呼ぶ）．同一の車両について，回答値と計算値が把握できた車両は1167台である．

回答値および計算値の分布を**図-1**，**図-2**にそれぞれ示す．図より，回答値の分布は，ところどころで極端に台数が多い区間が存在していることが分かる．特に，10000km～11000km，15000km～16000km 等の区間で台数が多くなっている．これらは，heaping と呼ばれる丸め誤差が回答に生じていることを示しているものである．一方，計算値の分布は，若干の出入りはあるものの，回答値に比べてなめらかな分布となっており，対数正規分布に近い形状であることが分かる．

本研究では，回答値を通常の調査と同様の精度で得られるデータとみなし，計算値を真値とみなして分析を行う．計算値にも，各調査時点でオドメーターを読み取る際に丸め誤差が発生する可能性がある

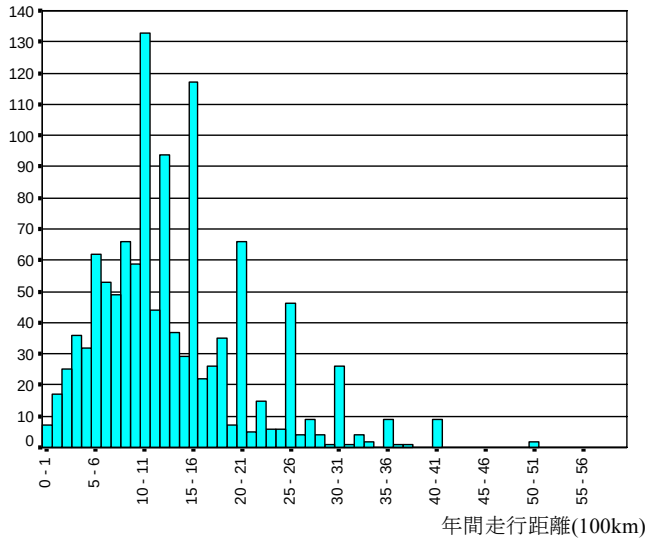

図-1　被験者の回答値の分布

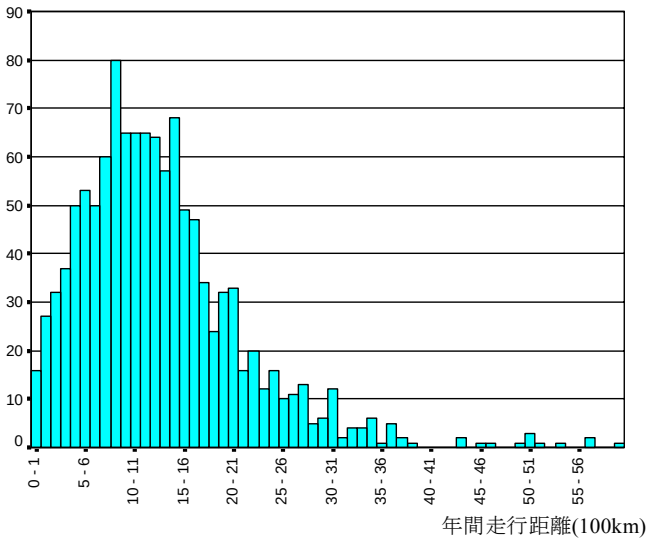

図-2　パネル調査による計算値の分布

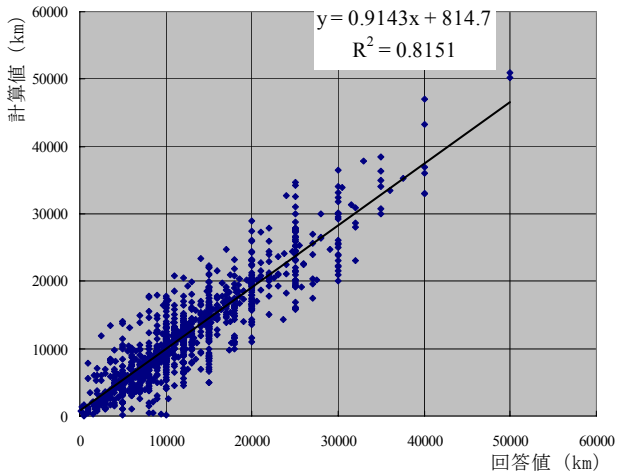

図-3　計算値と回答値の分布

が，**図-1**，**図-2**の結果からも，回答値に比べてその影響は小さいと考えられる．

次に，回答値が計算値からどのように乖離しているか，同一車両の回答値と計算値の散布図を図-3に示す．ただし，回答値と計算値の乖離が非常に大きいいくつかのケースについては，1997年の保有車両と1998年の保有車両をマッチングする時点で問題があったものと考え削除している．図より，全体として，回答値は計算値より若干大きな値をとっているものの両者の相関は高く，回答値はそれほど偏ってはいないと考えられる．ただし，回答値のいくつかの値でプロットが縦に並んでおり，回答値が丸め誤差を含んでいることが確認できる．

3.　報告誤差の分析

ここでは，個々の報告誤差はどのような場合に大きくなるかを検討するために，計算値と回答値の差に着目した回帰分析を行う．被説明変数は，計算値と回答値の差の絶対値，回答値が計算値より小さい場合（過小報告）の回答値と計算値の差，回答値が計算値より大きい場合（過大報告）の回答値と計算値の差，のそれぞれ対数をとった値を被説明変数とした回帰モデルを構築した．モデルの推定結果を**表-1** に示す．

表より，絶対値，過小報告ともに，保有台数の係数が負で有意な値を示している．また，過大報告についても有意ではないものの，負の値を示している．これより，保有台数が多い世帯については，報告誤差が小さいことが分かる．一方，通勤使用ダミーについては，過小報告について正に有意な値を示しており，通勤トリップへの自動車利用が認識以上の走行距離をもたらしている可能性を示唆している．また，絶対値については，ディーゼル車ダミーが正の，小型車ダミーが負の，それぞれ有意な値を示している．後述する年間走行距離モデルの推定結果より，ディーゼル車は年間走行距離が長く，小型車は年間走行距離が短くなる傾向が確認されており，これらの結果は，年間走行距離が長いほど報告誤差が大きくなることを示しているものと考えられる．絶対値のモデルについては，別途，年間走行距離の計算値を説明変数に追加したモデルを構築したところ，有意ではないものの，年間走行距離の計算値が正の値を持つ一方で，ディーゼル車および小型車も有意な値を取らなかった．これらの結果は，年間走行距離

表-1　報告誤差モデルの推定結果

変数名	絶対値推定値	過小報告推定値	過大報告推定値
定数項	7.513	7.781	7.254
15 才以下人数	0.033	0.075	-0.037
公共交通近在ダミー	0.126	-0.025	0.245
大都市在住ダミー	-0.007	0.212	-0.091
保有台数	-0.225**	-0.376**	-0.126
低所得層ダミー	-0.037	0.423	-0.379
高所得層ダミー	0.002	-0.169	0.149
39 歳以下ダミー	-0.044	-0.188	0.089
60 歳以上ダミー	-0.142	-0.252	-0.108
雇用者ダミー	-0.020	-0.169	0.121
男性ダミー	-0.079	-0.115	-0.026
通勤使用ダミー	0.163	0.355*	0.001
ディーゼル車ダミー	0.192*	0.137	0.243
小型車ダミー	-0.204*	-0.261	-0.193
大型車ダミー	0.105	-0.005	0.133
トラックダミー	0.129	1.266	-0.437
車齢	-0.018	-0.028	-0.006
サンプル数	965	433	532
決定係数	0.010	0.020	0.013

*10%有意, 5%有意

が長いほど報告誤差が大きくなるという推察を支持するものと考えられる.

4.　報告誤差を考慮した年間走行距離モデル

通常の年間走行距離の分析では，年間走行距離の対数を被説明変数とした回帰モデルが適用される場合が多い．回帰モデルは以下の式で表される.

$$y_i = \boldsymbol{\beta}\mathbf{x}_i + \varepsilon_i \tag{1}$$

ただし，y_i は年間走行距離の対数，$\boldsymbol{\beta}$は未知パラメータベクトル，$\mathbf{x}_i$ は説明変数ベクトル，ε_i は正規分布に従う誤差項.

通常の回帰分析では，誤差は正規分布に従うと仮定されているが，2 章で確認したように，年間走行距離の回答値は丸め誤差を含んでおり，誤差が正規分布とはなっていない．さらに，年間走行距離が長いほど丸め誤差が大きくなっているため，丸め誤差を無視した場合には推定値にバイアスが生じることが指摘されている [1].

ここで，丸め誤差の大きさが明らかな場合，オーダードプロビットモデルの適用が考えられる．オーダードプロビットモデルでは，先験的に決定した丸め誤差の大きさに基づき，例えば 1000km 単位に丸められると仮定した場合には, 以下の式が得られる.

$$\begin{aligned} y_i' &= 0 \quad \text{if } y_i < 500, \\ &= 1000 \quad \text{if } 500 \le y_i < 1500, \\ &= 2000 \quad \text{if } 1500 \le y_i < 2500, \\ &\vdots \end{aligned} \tag{2}$$

$$\begin{aligned} \Pr(y_i') &= \int_{\mu_{i-1}-\boldsymbol{\beta}\mathbf{x}_i}^{\mu_i-\boldsymbol{\beta}\mathbf{x}_i} \frac{1}{\sigma_\varepsilon}\phi\left(\frac{\varepsilon_i}{\sigma_\varepsilon}\right)d\varepsilon_i \\ &= \Phi\left(\frac{\mu_i-\boldsymbol{\beta}\mathbf{x}_i}{\sigma_\varepsilon}\right) - \Phi\left(\frac{\mu_{i-1}-\boldsymbol{\beta}\mathbf{x}_i}{\sigma_\varepsilon}\right) \end{aligned} \tag{3}$$

ただし, y'_i は丸め誤差を含む年間走行距離の回答値, ϕ, Φ は標準正規確率密度関数，累積分布関数，σ_ε は誤差項の分散，μ_i，$\mu_{i\text{-}1}$ はそれぞれケース i の丸め誤差の上限値および下限値.

ここで，実際の年間走行距離データには様々な大きさの丸め誤差が含まれており，全ての被験者が同一の丸め誤差に従うとは考えられない.

様々な丸め誤差がデータに含まれている場合の取り扱いについて，Heitjan & Rubin (1990)[2]はアフリカにおける年齢調査を対象として潜在クラスモデルを適用した分析を行っている. 本研究では, Heitjan & Rubin で用いられた方法を年間走行距離モデルに適用することによって，ランダムな丸め誤差を考慮した分析を行う.

本研究は，丸め誤差の大きさについて，基礎分析の結果より，500km 単位，1000km 単位，5000km 単位の 3 つの大きさの丸め誤差が混在しているものとする. 丸め誤差の大きさには序列性が存在するため, 個々の年間走行距離の回答値がいずれの丸め誤差に属するかを以下の式によって表す.

$$z_i^* = \alpha y_i + \boldsymbol{\Gamma}\mathbf{x}_i + \zeta_i \tag{4}$$

$$\begin{aligned} z_i &= 1 \quad \text{if } z_i^* < 0, \\ &= 2 \quad \text{if } 0 \le z_i^* < \theta, \\ &= 3 \quad \text{if } \theta \le z_i^* \end{aligned} \tag{5}$$

ただし，$\alpha, \boldsymbol{\Gamma}$ は未知パラメータおよび未知パラメータベクトル，ζ_i は正規分布に従う誤差項，θは未知の閾値，z_i = 1, 2, 3 はそれぞれ丸め誤差が 500km 単位，1000km 単位，5000km 単位のセグメントに属することを示す.

式(4)には丸め誤差を受ける前の年間走行距離の対数である y_i が含まれており，年間走行距離が長い

ほど丸め誤差が大きくなる可能性を考慮している．

ε_i と ζ_i は独立であると仮定すると，z_i^* と y_i の誤差の共分散行列は以下の行列で表される．

$$V\begin{pmatrix} z_i^* \\ y_i \end{pmatrix} = \begin{pmatrix} \sigma_\zeta^2 + \alpha^2\sigma_\varepsilon^2 & \alpha\sigma_\varepsilon^2 \\ \alpha\sigma_\varepsilon^2 & \sigma_\varepsilon^2 \end{pmatrix} \quad (6)$$

ここで，一般性を失うことなく $\sigma_\zeta^2 + \alpha\sigma_\varepsilon^2$ を 1 に基準化すると，z_i^* と y_i の誤差の相関は $\alpha\sigma_\varepsilon$ で表される．また尤度関数は以下の式で表される．

$$L = \prod_i \sum_{z \in Z_i} \Pr(y_i' | z) \Pr(z) \quad (7)$$

ただし，Z_i はケース i が帰属する可能性のある丸め誤差のセグメントの集合．式(7)では，13000km と答えている場合は 5000km 単位の丸め誤差ではありえないといった回答値から判断できる各セグメントへの帰属可能性の有無を考慮した潜在セグメントモデルとなっている．

モデルの推定結果を**表-2** に示す．本研究で提案したモデルの推定結果より，セグメント帰属関数について，α の推定値が有意に正の値を取っており，年間走行距離が長いほど丸め誤差が大きくなることが分かる．また大型車ダミーも有意に正の値を取っており，大型車の走行距離に関する報告に誤差が大きいことが示された．

一方，年間走行距離の説明変数については，同時に推定した，計算値を被説明変数とした回帰モデルおよび回答値を被説明変数とした回帰モデルと明確な差は認められず，本提案モデルによる説明力の向上は確認できなかった．ただし，年間走行距離モデルの誤差の標準偏差については，回答値を被説明変数とした回帰モデルより提案モデルの方が小さな値を取っており，全体として説明力が向上しているとも解釈できる．ただし，計算値を被説明変数とした回帰モデルでの誤差の標準偏差は他のどのモデルよりも大きいため，解釈には注意が必要と考えられる．

表-2　年間走行距離モデルの推定結果

モデル	回帰モデル		回帰モデル		提案モデル	
データ	計算値		回答値		回答値	
	推定値	t 値	推定値	t 値	推定値	t 値
定数項	9.380	70.92	9.345	81.53	9.340	83.11
15 才以下人数	-0.015	-0.45	-0.014	-0.49	-0.015	-0.43
公共交通近在ダミー	-0.070	-1.38	-0.061	-1.38	-0.048	-1.12
大都市在住ダミー	0.116	2.07	0.013	0.27	0.025	0.57
保有台数	-0.038	-0.89	0.017	0.47	0.014	0.36
低所得層ダミー	-0.115	-1.36	-0.202	-2.75	-0.220	-3.68
高所得層ダミー	0.041	0.80	0.057	1.29	0.051	1.25
39 歳以下ダミー	0.095	1.43	0.105	1.81	0.102	1.62
60 歳以上ダミー	-0.326	-3.82	-0.201	-2.72	-0.182	-2.37
雇用者ダミー	-0.137	-1.65	-0.074	-1.03	-0.073	-0.96
男性ダミー	0.115	2.32	0.110	2.56	0.095	2.18
通勤使用ダミー	0.390	6.30	0.357	6.65	0.354	6.45
ディーゼル車ダミー	0.389	8.17	0.379	9.17	0.372	8.59
小型車ダミー	-0.269	-5.28	-0.174	-3.93	-0.178	-4.05
大型車ダミー	0.163	2.09	0.160	2.36	0.151	2.45
トラックダミー	0.536	1.45	0.579	1.81	0.559	1.29
車齢	-0.037	-6.56	-0.038	-7.86	-0.037	-8.39
σ_ε	0.634		0.550		0.508	
セグメント帰属関数						
定数項					-8.166	-9.23
ディーゼル車ダミー					-0.093	-0.87
大型車ダミー					0.344	1.98
車齢					-0.010	-0.71
α					0.903	9.74
θ					0.851	10.97
サンプル数	975		975		975	
決定係数	0.324		0.347			
最終尤度					-3350.3	

5.　おわりに

本研究では，年間走行距離の報告における報告誤差の分析，および報告誤差を考慮した年間走行距離モデルを構築した．今後は，個人内での丸め誤差の時点間安定性の分析や，トリップ時間等を対象とした分析に取り組む予定である．

参考文献

1) Heitjan, D.F. and Rubin, D.B.: Ignorability and coarse data, *The Annals of Statistics*, Vol. 19, No. 4, pp. 2244-2253, 1991.
2) Heitjan, D.F. and Rubin, D.B.: Inference from coarse data via multiple imputation with application to age heaping, *Journal of the American Statistical Association*, Vol. 85, No. 410, pp. 304-314, 1990.