データの収集方法と精度に対応した分析手法

名古屋大学　山本俊行

NU TREND
Nagoya University
TRansport and ENvironment Dynamics

講義内容

- データと分析手法の関係
- 標本抽出が偏っている時
  - 偏りが既知のケース
  - 偏りが未知のケース
- 変数値の観測が不完全な時
  - 説明変数の不完全観測のケース
  - 被説明変数の不完全観測のケース

データと分析手法の関係

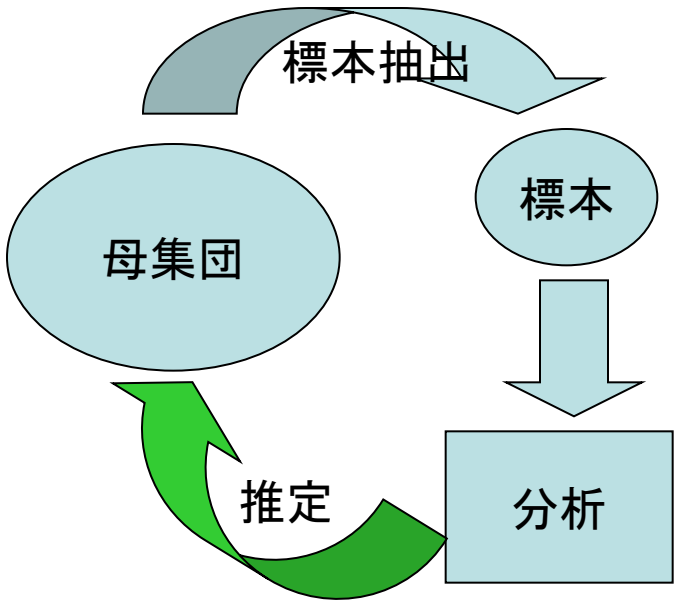
- 実務家にとっては、データを取って如何に検証したい内容を明確に日の下に晒せるかこそが腕の見せ所であって、複雑な方法論を使わないと「わからないデータ」を取ってしまった時点で、実は半分自分の無能を曝しているのと同義でないかと思うことがある。
- 実務家の視点で言えば、いくら手法がよくてもデータがダメではね、ということになる。

（行動計量学会報第100号(2004)での実務家の発言）

分析の目的と方法

目的:母集団中のパラメータを知る

方法:母集団から標本を抽出し標本を調べる



- 標本抽出と標本の観測を工夫し，仮説を簡単な分析で明らかにする
  - 実験計画等
- 標本抽出と標本の観測が不完全な時
  - 分析時に適切な対応が必要

離散選択モデルにおける標本抽出方法と推定方法

- 無作為抽出：標本分布は母集団分布と一致すると仮定できる場合，通常の最尤推定量

$$\ln L = \sum_{i=1}^{N} \ln \Pr[J_i \mid \mathbf{x}_i, \boldsymbol{\theta}]$$

- 選択肢別抽出：選択肢毎に抽出率が異なり，抽出率が既知の場合，WESML推定量

$$\ln L = \sum_{i=1}^{N} \hat{\omega}(\mathbf{j}_i) \ln \Pr[J_i \mid \mathbf{x}_i, \boldsymbol{\theta}]$$

$$\hat{\omega}(\mathbf{j}_i) = \left[\frac{H(\mathbf{j}_i)}{Q(\mathbf{j}_i \mid \boldsymbol{\theta})} \right]^{-1}$$

H: 標本内の比率，Q: 母集団内の比率

- WESML推定量を用いた場合のパラメータ推定値の分散共分散行列は，サンドイッチ推定量で与えられる

$$\Sigma = \frac{1}{N}\Omega^{-1}\Lambda\Omega^{-1}$$

$$\hat{\Omega} = \frac{1}{N}\sum_{i=1}^{N}\left\{\left[\frac{\partial \ln \Pr(J_i \mid \mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]\left[\frac{\partial \ln \Pr(J_i \mid \mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right]\right\}$$

$$\hat{\Lambda} = \frac{1}{N}\sum_{i=1}^{N}\left\{\hat{\omega}(\mathbf{j}_i)\left[\frac{\partial \ln \Pr(J_i \mid \mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]\left[\frac{\partial \ln \Pr(J_i \mid \mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right]\right\}$$

通常の統計パッケージソフトでは，重み付き最尤推定を計算しただけでは分散共分散行列が Λ で計算される場合も多いので注意！

多次元選択肢別抽出法への応用

- 観光行動調査など，観光目的地，宿泊地，駅や高速道路IC等の様々な場所で調査票が配布される
- 配布場所毎に抽出率は異なってしまう

方法

- 特定の個体が標本として複数次元（観光目的地，宿泊地，駅や高速道路IC等の組み合わせ選択）で抽出される確率を無視する
- 各次元に対して計算される重みを統合した重みを用いる

$$\omega(\mathbf{j}) = \left[\sum_{d} \sum_{b, j_d \in C_b^d} \frac{H^d(b)}{Q^d(b \mid \boldsymbol{\theta})} \right]^{-1}$$

d: 次元

- 分析対象の離散選択モデルの選択肢と一致するわけではない

交通行動の分析とモデリング(2002)を参照のこと

抽出率が未知の場合はどうなるのか

Cosslett（1981)の提案した推定量

$$\ln L = \frac{1}{N}\sum_{i=1}^{N} \ln \frac{P(y_i = j_i \mid \theta) H(j_i) / Q(j_i)}{\sum_{k=1}^{J} P(y_i = k \mid \theta) H(k) / Q(k)}$$

θ に加えて Q も未知パラメータとして推定する

- そんな場合はあるのか？
- 実務家が無能と呼ぶ場合なのか？

Cosslett (1981) Maximum likelihood estimator for choice-based samples, Econometrica, Vol. 49
Cosslett (1981) Efficient estimation of discrete-choice models, Structural Analysis of Discrete Data with Econometric Applications (Manski & McFadden eds.)

身体損傷程度が記録された交通事故データ

- 損傷程度が低いほど報告されない確率が高い
- 報告されない事故の数は分からない

米国ワシントン州の路側障害物衝突事故データ

	都市部		郊外部	
	サンプル数	%	サンプル数	%
車両損傷のみ	6125	63	6514	61
損傷の可能性有り	1646	17	1357	13
軽傷	1602	16	2191	21
重傷	297	3	474	4
死亡	53	1	104	1
合計	9723	100	10640	100

オーダードプロビットモデルを用いた推定結果

報告漏れ推定結果（上段：事故数推定値，下段：報告漏れ率）

	都市部				郊外部			
報告漏れの存在	0	0,1	0,1,2	報告数	0	0,1	0,1,2	報告数
0：車両損傷のみ	4298 -42.5%	8349 26.6%	2824 -116.9%	6125	3447 -89.0%	3237 -101.2%	4544 -43.4%	6514
1：損傷の可能性有り		**6867** **76.0%**	**3049** **46.0%**	1646		**2522** **46.2%**	**3268** **58.5%**	1357
2：軽傷			905 -77.0%	1602			2704 19.0%	2191
3：重傷				297				474
4：死亡				53				104
合計	7896 -23.1%	17168 43.4%	7128 -36.4%	9723	7573 -40.5%	8528 -24.8%	11094 4.1%	10640

太字：5%有意

分析結果

- 報告漏れを多くの損傷程度に仮定すると推定が困難となった
- ほとんどの説明変数のパラメータ推定値は報告漏れを考慮してもほとんど変化しなかった
- 「損傷の可能性あり」の事故の報告漏れのみが有意となった
  - 損傷の可能性あり，というカテゴリーが警察官の恣意的な判断が含まれる？
  - 「車両損傷のみ」は最小のカテゴリーであり，より軽度な車両損傷事故は全く報告されない場合，「報告されるレベルの車両損傷」というカテゴリーに属する事故の報告漏れは小さい？

変数値の観測が不完全の場合：不完全データ

- 欠測：変数値が得られないケース

- Coarse：観測が正確でないケース
  - Censoring：一定以上の値であることしか分からない
  - Rounding:整数値等に丸められる
  - Heaping：さまざまなレベルのRoundingが含まれる

説明変数の観測が正確でない例

- コンパクトシティを交通行動から評価する際，駅からの距離等の立地条件を詳細に表現することが必要
- PT等の一般的な交通調査で用いられるゾーンシステムは詳細な立地条件を十分表現できない

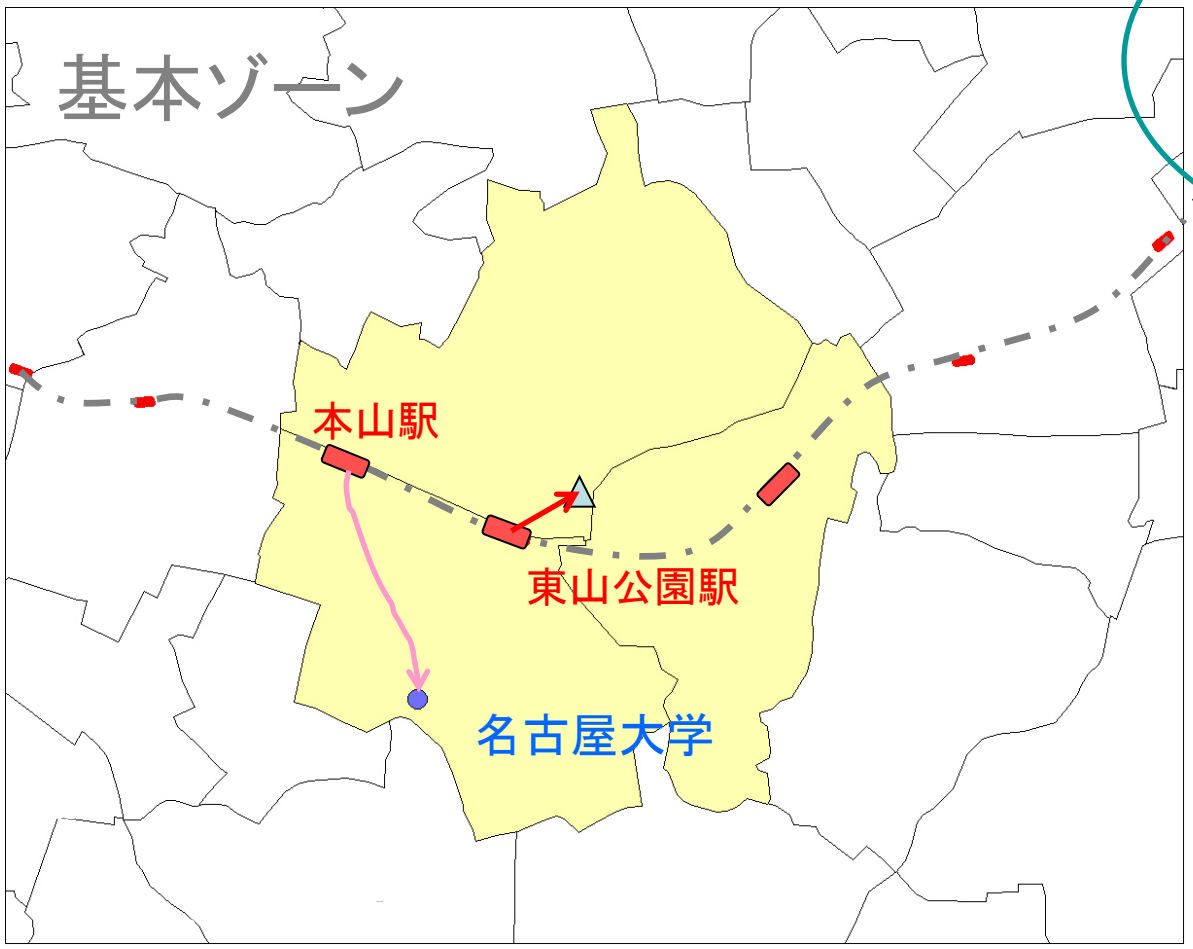
分析目的に対して，駅アクセス距離の観測が正確でない

駅アクセス距離の算出

実測　⇒730m

小ゾーン　⇒490m

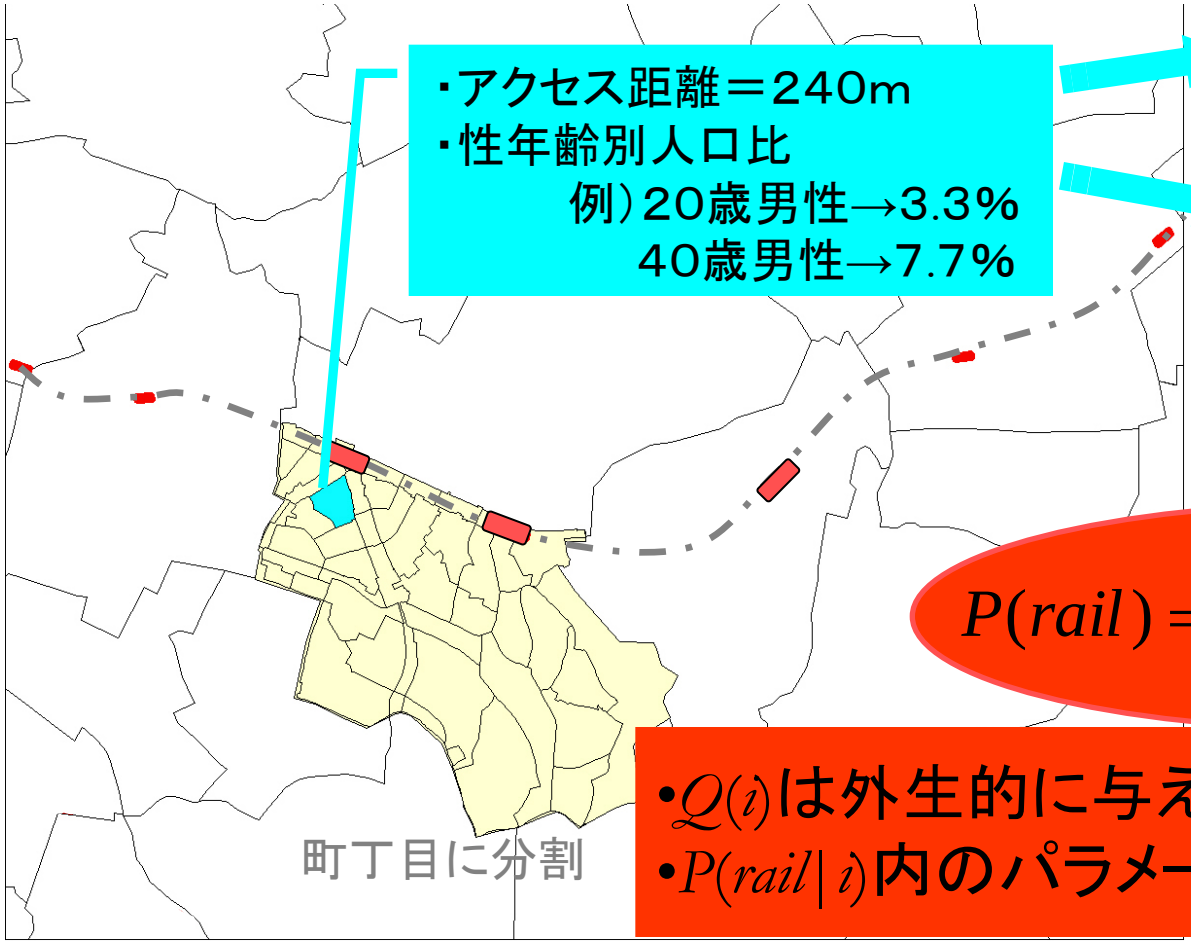
基本ゾーン⇒330m



誤差が生じる

アクセス距離の不正確さを考慮した推定法

出発点が町丁目iにある時の、



・アクセス距離＝240m
・性年齢別人口比
例）20歳男性→3.3%
40歳男性→7.7%

$P(rail \mid i)$ 鉄道の選択確率

$Q(i)$ 存在確率

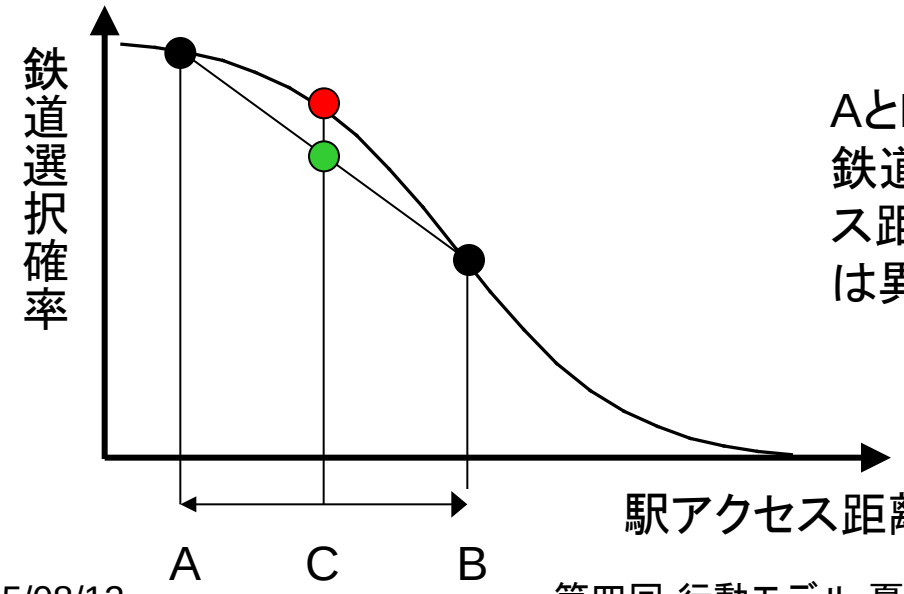
潜在クラスモデル

$$P(rail) = \sum_i \{Q(i) \times P(rail \mid i)\}$$

- $Q(i)$は外生的に与えられる
- $P(rail \mid i)$内のパラメータは i によらず共通

注意

- 離散選択モデルは非線形なため，前ページで説明した潜在クラスモデルと，ゾーン内で存在確率を用いた重み付き平均アクセス距離を用いたモデルとは一致しない
- 平均的個人の誤謬と同じ構造



AとBに居住している確率が50%の時，鉄道選択確率の期待値●は，平均アクセス距離を持つC地点の鉄道選択確率●とは異なる

交通手段選択モデルの
パラメータ推定結果（括弧内t値）

	潜在クラス	最小ゾーン	基本ゾーン
代表手段			
バス停アクセス	-9.2 (-5.6)	-2.5 (-8.8)	-2.5 (-9.3)
鉄道端末			
バス停アクセス	-1.4 (-4.0)	-1.2 (-4.6)	-1.1 (-4.7)
駅アクセス	-1.4 (-11.9)	-0.8 (-11.8)	-0.7 (-11.0)
最終尤度	-2900	-2944	-2944

潜在クラスモデルの適用により

- 適合度が向上
- パラメータ値が増大

被説明変数がheap している例

- 自動車の年間走行距離を被験者に質問する場合，被験者の記憶力に限界があるため正確な走行距離が観測できない
- 被験者は，きりの良い値で答えるが，得られたデータだけからは，どの単位で丸められたか判断できないケースがある

フランスParc-Autoデータ

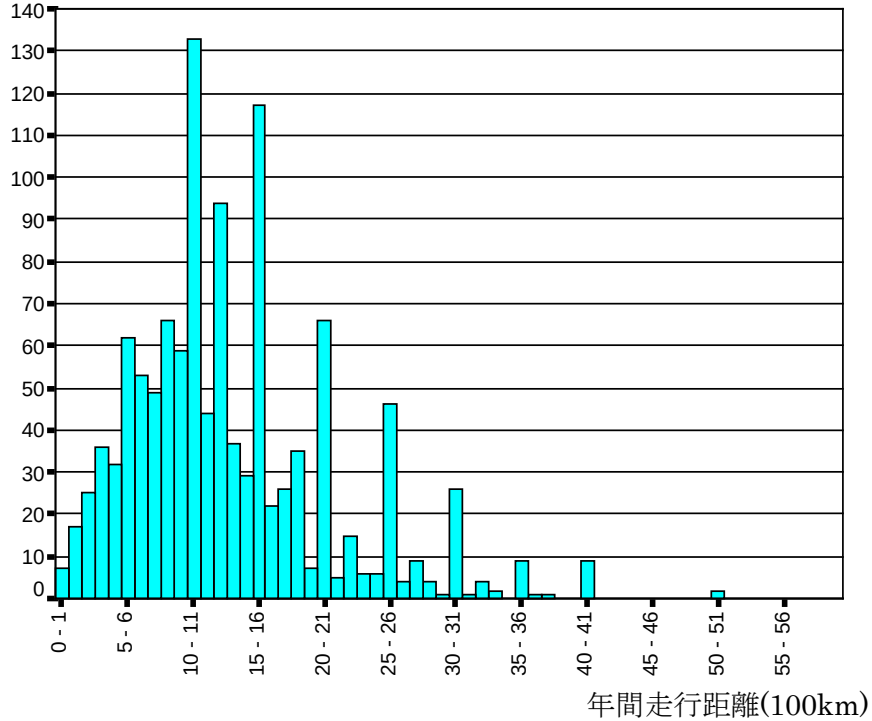


図-1　被験者の回答値の分布

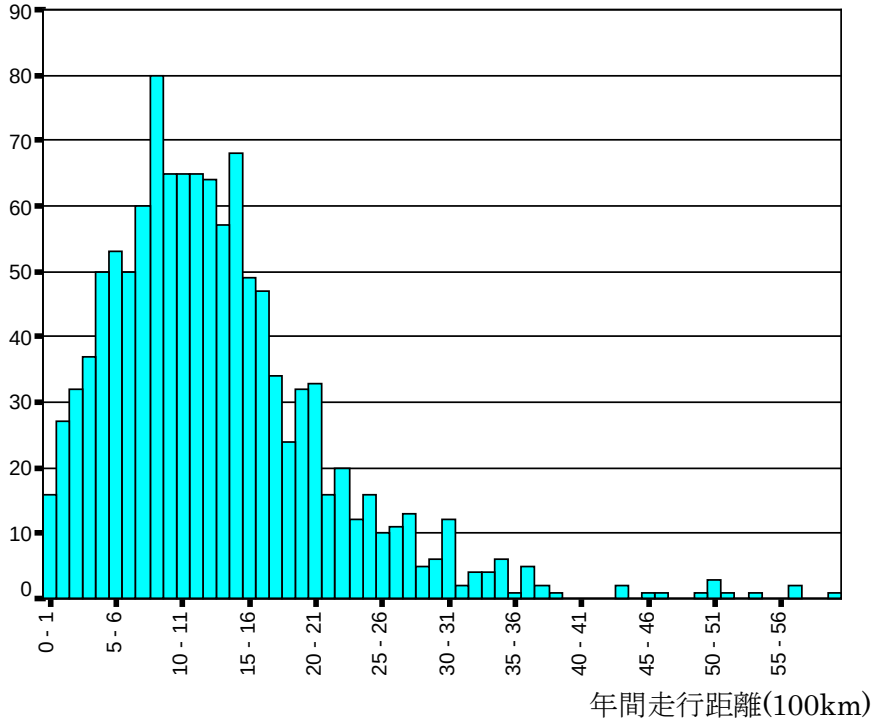
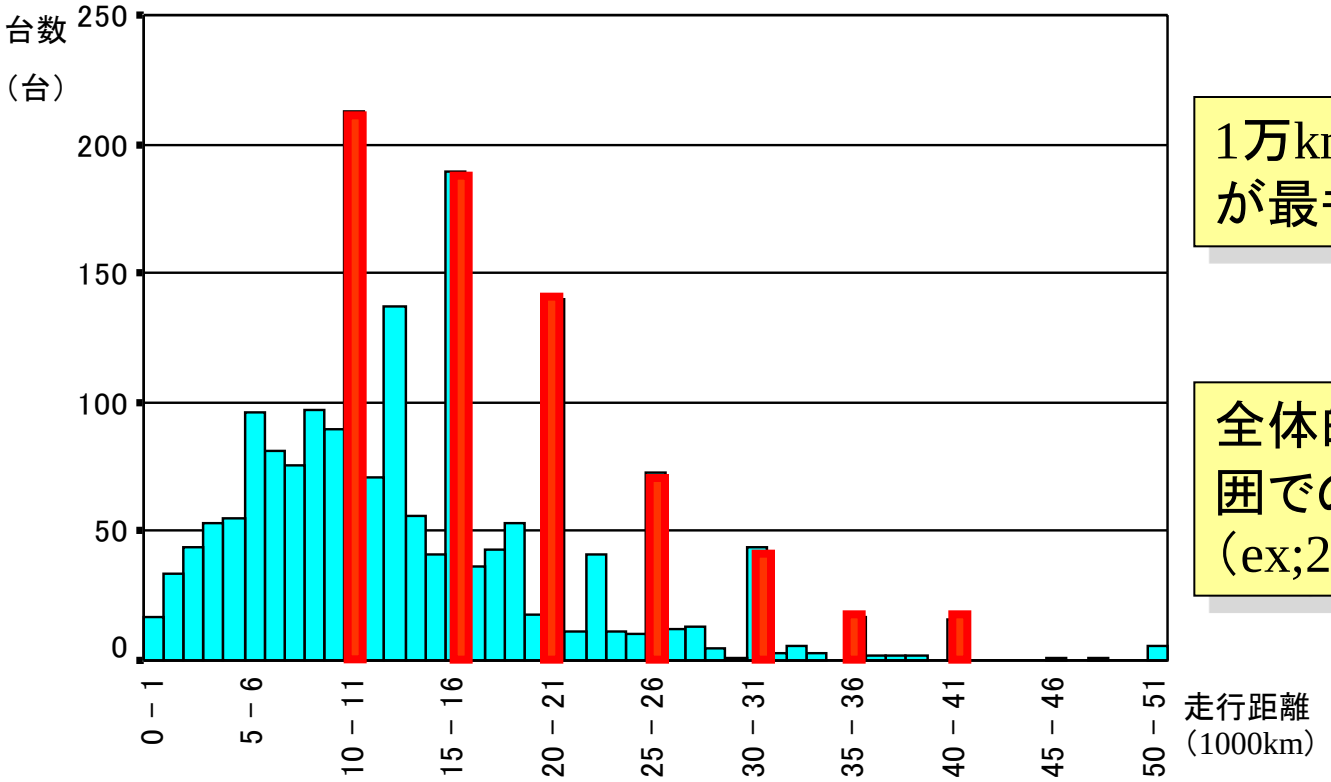


図-2　パネル調査による計算値の分布

(オドメータ値の増加量)

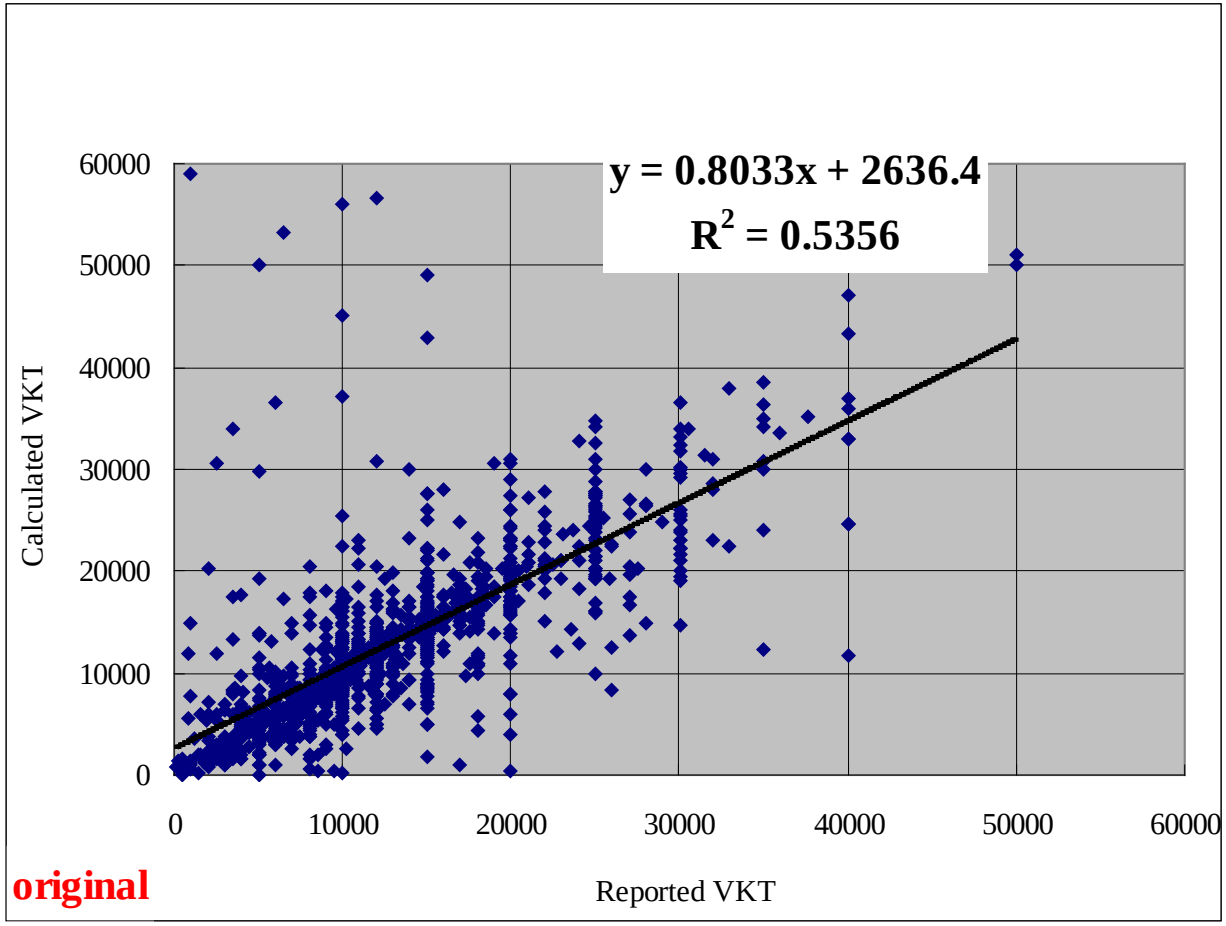


年間走行距離の分布（1998年）　（N=1919）

1万kmを含む範囲の回答が最も多い.

全体的に5千kmごとの範囲での回答が多い.
（ex;2.5万km, 3万km）

回答値と計算値の関係



回答の誤記入やデータ処理のミスが疑われるケースもあるものの，例えば

- 回答値20000kmの場合の計算値のばらつきは，15000kmや25000kmの場合よりも大きい傾向がある

分析方法

- 年間走行距離モデルは回帰モデルで表す

$$y_i^* = \beta \mathbf{x}_i + \varepsilon_i$$

- 丸め誤差は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し，オーダードプロビットモデルで表す．また，年間走行距離そのものが丸め誤差に影響を与える

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

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

- 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

- 年間走行距離の観測値は，例えば500km単位で丸められている場合，以下で表される

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

- 丸め誤差の単位が不明であるため，尤度は以下の潜在クラスモデルで与えられる

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

- ただし，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

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

推定結果(一部)

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
走行距離モデルの誤差項の標準偏差	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	

分析結果

- 年間走行距離が長いほど丸め誤差が大きくなる
- 大型車の走行距離に関する報告に誤差が大きい
- 説明変数の推定値にはそれほど影響がない

今後の課題

- 個人内での丸め誤差の時点間安定性の分析や，トリップ時間等の分析への適用

最後に

- やはりデータを取る時点で良いデータが得られるように努力すべき
- それでも完全なデータは得られない事が多い
- 不完全なデータでも分析手法を工夫することで，よりよい知見を得られる可能性がある