交通行動分析における
不完全データの取り扱い

名古屋大学　山本俊行

NU TREND
Nagoya University
TRansport and ENvironment Dynamics

講演内容

- 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（ケース１）
  - 偏りが未知のケース（ケース２）
- 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（ケース３）
  - 被説明変数の不完全観測のケース（ケース４）

データと分析手法の関係

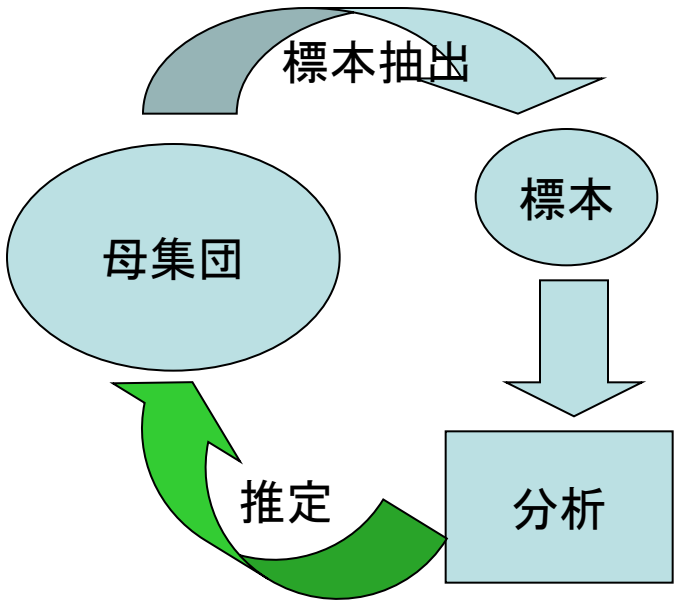
- 実務家にとっては、データを取って如何に検証したい内容を明確に日の下に晒せるかこそが腕の見せ所であって、複雑な方法論を使わないと「わからないデータ」を取ってしまった時点で、実は半分自分の無能を曝しているのと同義でないかと思うことがある。
- 実務家の視点で言えば、いくら手法がよくてもデータがダメではね、ということになる。

（芳賀麻誉美：学会への提言，行動計量学会報第100号，pp. 1-2, 2004）

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

標本抽出方法による尤度の違い

- 無作為抽出: $\lambda = \Pr[j_i \mid \mathbf{x}_i, \boldsymbol{\theta}] p(\mathbf{x}_i)$ より

$$\ln \lambda = \ln \Pr[j_i \mid \mathbf{x}_i, \boldsymbol{\theta}] + \ln p(\mathbf{x}_i)$$

  - 第二項には θ が含まれていないため省略可能

- 外生層別抽出: $\lambda = \Pr[j_i \mid \mathbf{x}_i, \boldsymbol{\theta}] \frac{p(\mathbf{x}_i)}{\int_{\mathbf{z_b}} p(\mathbf{y}) d\mathbf{y}} H(\mathbf{b})$ より

$$\ln \lambda = \ln \Pr[j_i \mid \mathbf{x}_i, \boldsymbol{\theta}] + \ln p(\mathbf{x}_i) - \ln \int_{\mathbf{z_b}} p(\mathbf{y}) d\mathbf{y} + \ln H(\mathbf{b})$$

  - 同じく第二項以降は省略可能

Pr[]: 行動モデル, *j*: 選択肢, **x, y, b**: 外生変数, θ: パラメータ, *p*(): 外生変数の確率分布, H(): 抽出層分布

- 選択肢別抽出 $\lambda = \Pr[j_i \mid \mathbf{x}_i, \boldsymbol{\theta}] \frac{p(\mathbf{x}_i)}{Q(j_i)} H(j_i)$ ただし

$Q(j_i) = \int_{j_i \times \mathbf{z}} \Pr[j_i \mid \mathbf{y}, \boldsymbol{\theta}] p(\mathbf{y}) d\mathbf{y}$ より,

$\ln \lambda = \ln \Pr[j_i \mid \mathbf{x}_i, \boldsymbol{\theta}] + \ln p(\mathbf{x}_i) - \ln Q(j_i) + \ln H(j_i)$

  - Q には θ が含まれるため無視できない
  - Q が既知のとき θ の制約条件式になる

通常の最尤推定法の式は用いることが出来ない

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

ケース1：
複雑な内生抽出法に基づく標本への離散選択モデルの適用

北村隆一，酒井弘，山本俊行：複雑な内生抽出法に基づく標本への離散選択モデルの適用，土木学会論文集，No. 667/IV-50, pp. 103-111, 2001.

背景

観光行動調査によって得られたデータは京都市の観光客全体を代表しているのか？

- 対象となる調査では，幅広くサンプルを得るために，様々な地点で調査票を配布している.
  - 全ての観光地点では配布できない.
  - 地点によって配布数が異なる.

京都市観光行動調査

- 調査実施日時：1997年11月3日（秋の観光シーズンのピーク）
- 調査内容：全ての目的地および滞在時間，個人属性，当日の来洛体験についての態度
- 調査規模：総計26,688部の調査票を配布，5,692が郵送回収．返答率21.3%
  - 調査票が色刷りで見やすい
  - 抽選で回答者500名に京舞妓の写真入りの「オリジナル・テレフォンカード」を贈呈
  - 調査が対象とする観光行動が回顧するに楽しいものである

調査票配布地点

- 観光地（清水寺など23ヶ所）
- 駅（京都駅・観光1日乗車券発売所など15ヶ所）
- インターチェンジ（名神京都南 I.C.など2ヶ所）
- 宿泊施設66ヶ所
- 都心地域（四条大橋・河原町三条）

観光地の入り込み客数
（事前調査結果）

観光地	入り込み客数*	観光地来訪比率
清水寺	20,106人	26.7%
嵐山観月橋	15,510人	19.7%
金閣寺	7,260人	14.9%
銀閣寺	8,270人	14.5%

*10時から17時

鉄道駅での抽出率

鉄道駅	乗車客数*	回収数	抽出率
京都駅(JR東海)	42,792	130	0.30%
京都駅(JR西日本)	180,201	294	0.16%
二条駅(JR西日本)	6,100	76	1.25%
出町柳駅(京阪電車)	42,032	78	0.19%
烏丸駅(阪急電車)	61,246	132	0.22%
大宮駅(京福電車)	9,332	157	1.68%

*実測値は観測されていないため推計値

名神インターにおける抽出率

インター名	乗用車(台)*	回収数	抽出率
京都東	11,468	321	2.80%
京都南	20,385	488	2.39%

＊京都東～滋賀県境，京都南～京都東，京都市境～京都南の区間での乗用車比率をインター利用総台数に乗じたもの

宿泊施設の抽出率

宿泊施設	集客数*	回収数	抽出率
ホテル	13,152	507	3.86%
公的宿泊施設	1,616	144	8.91%
旅館	2,134	113	5.30%
ペンション	480	50	10.42%

＊配布対象の宿泊施設の部屋数×宿泊人数

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

入洛交通機関の選択モデル構築に際し，重みを計算する必要があるが，抽出は利用交通機関に相関があるものの，純粋に交通機関別でもない

対応方法

- 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

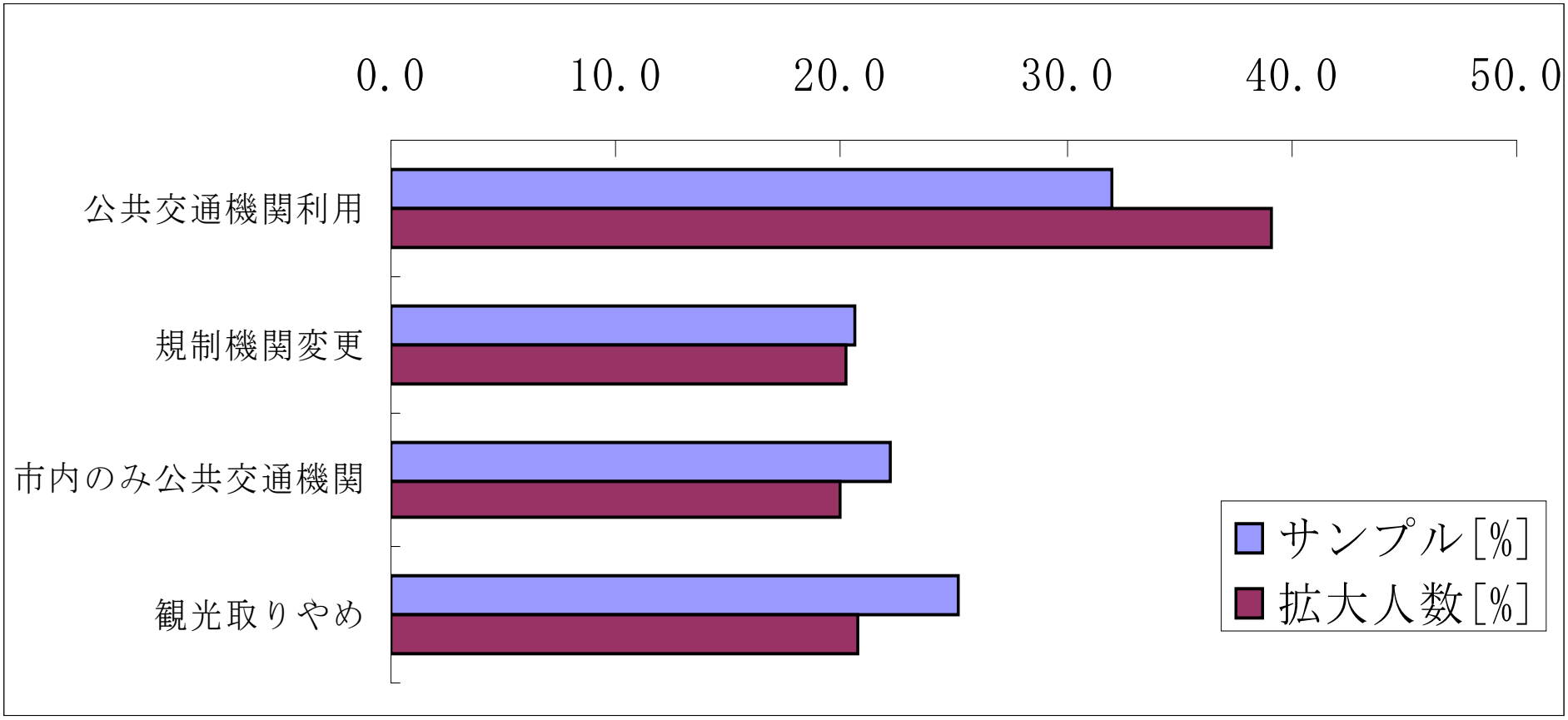
$$\omega(\mathbf{j}) = \left[\sum_{d} \sum_{b, j_d \in C_b^d} \frac{H^d(b)}{Q^d(b \mid \boldsymbol{\theta})} \right]^{-1}$$

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

サンプルの拡大の有無による違い（自動車規制時の対応行動）



自動車利用選択モデル

説明変数	重みなし		提案法	
	β	t	β	t
男性	0.64	7.59	0.56	6.38
50歳以上	-0.61	-7.47	-0.84	-10.06
就業者	0.21	2.17	0.43	4.41
毎日車を運転	0.92	11.76	0.97	12.41
家族と訪洛	0.88	11.06	0.97	12.21
初訪洛	-0.41	-3.16	-0.55	-3.29
自動車所要時間	-1.11	-5.34	-0.65	-2.73
自動車費用	**-0.42**	**-0.90**	**-1.71**	**-3.10**
鉄道所要時間	0.96	6.11	0.91	5.02
鉄道費用	**0.49**	**2.32**	**1.08**	**4.23**
定数項	-1.88	-16.95	-2.40	-21.40

- ●交通手段別費用の係数の差異が特に顕著
- ●同様の差異がt-値の推定値についても見られる.

まとめ

- 複雑な内生標本抽出法により得られた標本に適用可能な一意的な重みが存在する
- 多次元選択肢別抽出に対応した重みを設定することで，パラメータの推定バイアスを補正することが出来る
- 複雑な内生標本抽出を実施する場合，調査票配布場所毎の母数の把握が重要

ケース2：
報告漏れを考慮した
交通事故データの分析

山本俊行，端地純平：報告漏れを考慮した交通事故データの解析，土木計画学研究・講演集，Vol. 32, CD-ROM, 2005.

選択肢別抽出で抽出率が未知の場合はどうするのか

Cosslett（1981)の提案した推定量

$$\ln L = \frac{1}{N}\sum_{i=1}^{N} \ln \frac{\Pr\left[j_i \middle| \mathbf{x}_i, \boldsymbol{\theta}\right] H(j_i)/Q(j_i)}{\sum_{k=1}^{K} \Pr\left[k \middle| \mathbf{x}_i, \boldsymbol{\theta}\right] H(k)/Q(k)}$$

$\boldsymbol{\theta}$ 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？
- 分析者が無能な場合なのか？

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.)

背景

- 効率的な交通安全施策を実施するには、交通事故データの分析を行い、事故の要因などを把握することが重要

しかし

- 損傷程度の軽い事故の場合は報告されない可能性があり、交通事故データに含まれていないことが考えられる

実際の交通事故の要因を分析できていない可能性がある

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

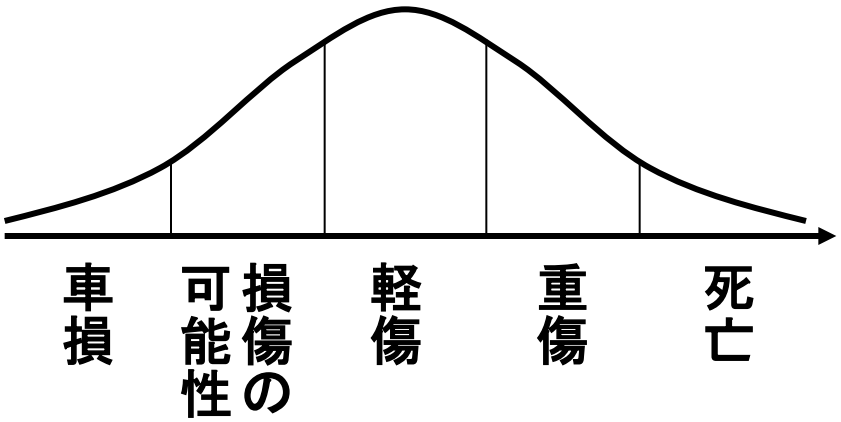
身体損傷程度のモデル

- Ordered probit Model

車両損傷のみ or 損傷の可能性有り

⋮

重傷 or 死亡



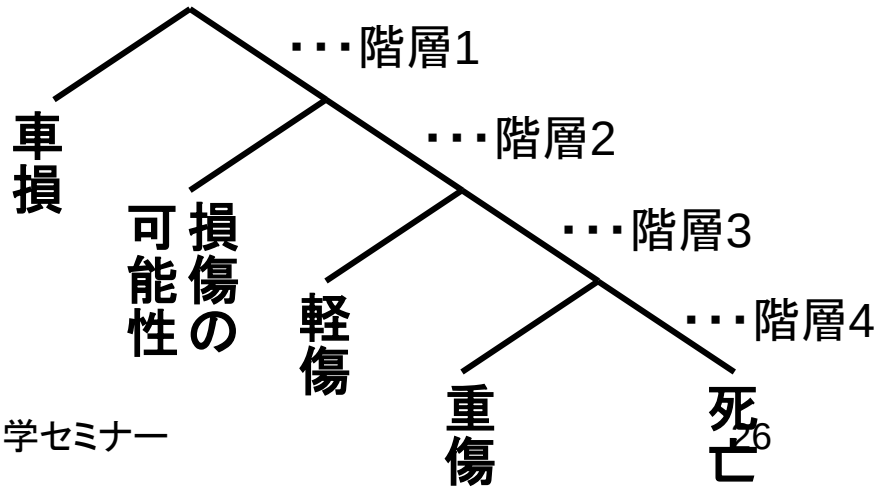
同じ関数で説明

- Sequential binary probit model

（階層1） 車両損傷のみ or 損傷の可能性有り以上

⋮

（階層4） 重傷 or 死亡



異なる関数で説明

分析の手順

1.報告漏れを考慮しない場合

Ordered probit model

Sequential binary probit model

比較

2.報告漏れの比率を仮定し、重み付けして推定

Ordered probit model

Sequential binary probit model

WESML法

3.報告漏れの比率が未知として推定

Ordered probit model

Sequential binary probit model

Cosslettの方法

報告漏れを考慮しない場合

- AICを用いた適合度の比較

$AIC = -\ln L(\beta^*) + K$

$L(\beta^*)$：最終尤度

K：未知パラメータ数

値が小さいほど適合度が良い

	AIC	
	都市部	郊外部
Ordered	9234.5	10337.6
Sequential	9145.8	10261.0

都市部・郊外部ともに、Sequential binary probit modelの方が適合度が良い

WESML法

- 報告漏れの影響を分析

パラメータを個別に比較

報告漏れを仮定したモデル（WESML法）の推定値 ⇔ t検定 ⇔ 報告漏れを考慮しないモデルの推定値

Ordered probit					
「車両損傷のみ」の報告漏れ率（%）	10	20	30	40	50
定数項					1
説明変数					
閾値1		1	1	1	1
閾値2		1	1	1	1
閾値3			1	1	1
合計	0	2	3	3	4

Sequential binary probit					
「車両損傷のみ」の報告漏れ率（%）	10	20	30	40	50
定数項				1	1
説明変数					
合計	0	0	0	1	1

・報告漏れ率が大きくなるほど推定値への影響が大きくなる傾向
・説明変数に有意差のあるものは無い

WESML法

- 報告漏れの影響を分析

パラメータをまとめて比較

報告漏れを仮定したモデル（WESML法）

← χ^2検定

報告漏れを考慮しないモデルの推定値

		χ^2値									
報告漏れ率（%）	車両損傷のみ	10	20	30	40	50	10	20	30	40	50
	損傷の可能性						5	10	15	20	25
Ordered probit		21.3	94.2	234.0	461.9	807.6	14.1	63.0	158.0	315.8	562.5
Sequential binary probit		0.01	94.7	234.6	462.5	808.2	2.1	8.9	21.1	39.7	66.0

・20～30%で報告漏れの影響

・報告漏れを「車両損傷のみ」に仮定した場合より「車両損傷のみ」と「損傷の可能性有り」に報告漏れを仮定したモデルの方が影響が小さい傾向

Cosslettの方法

- 事故件数推計値

報告漏れを仮定した損傷程度	Ordered			Sequential				報告された事故件数
	車損	車損･損可	車損～軽傷	車損	車損･損可	車損～軽傷	車損～重傷	
車両損傷のみ	4298	8349	2824	1240	22275	4228	32832	6125
損傷の可能性	1646	6867	3049	1646	62684	11897	92389	1646
軽傷	1602	1602	905	1602	1602	20	159	1602
重傷	297	297	297	297	297	297	2665	297
死亡	53	53	53	53	53	53	53	53
合計	7896	17168	7128	4838	86911	16496	128098	9723

網掛け箇所 ･･･発生事故件数＜報告事故件数

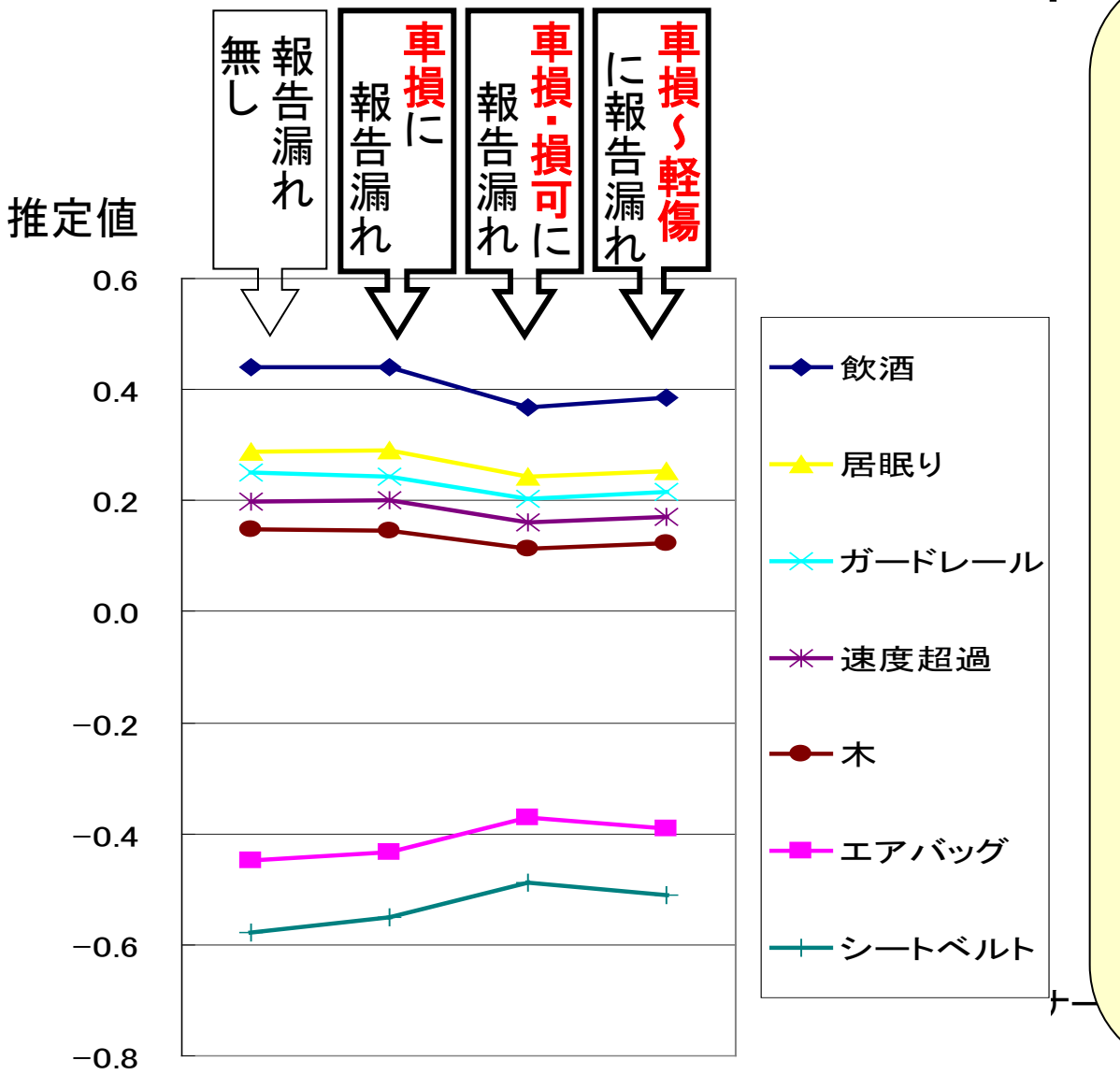
赤字 ･･･報告漏れが統計的に有意

･両モデルとも、「車両損傷のみ」と「損傷の可能性有り」に報告漏れを仮定した場合に矛盾の無い推定結果

･「損傷の可能性有り」のみ報告漏れが有意

Cosslettの方法

- パラメータ推定値（ordered probit）



特徴1

Cosslettの方法を用いた推定値

報告漏れを考慮しないモデルの推定値

特徴2

車両損傷のみ
損傷の可能性有り

車両損傷のみ
損傷の可能性有り
軽傷

車両損傷のみ

まとめ

- 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
- 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，というカテゴリーには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？

ケース3：
交通手段選択分析における
潜在クラスモデルによる
起終点位置観測精度の補完

山本俊行，小森陵補：交通手段選択分析における潜在クラスモデルによる起終点位置観測精度の補完，土木計画学研究・講演集，Vol. 30, CD-ROM, 2004.

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

使用するデータ①

中京都市圏PT調査データ

・サンプル数が非常に多い

・立地条件に関する詳細な記載がない

対策

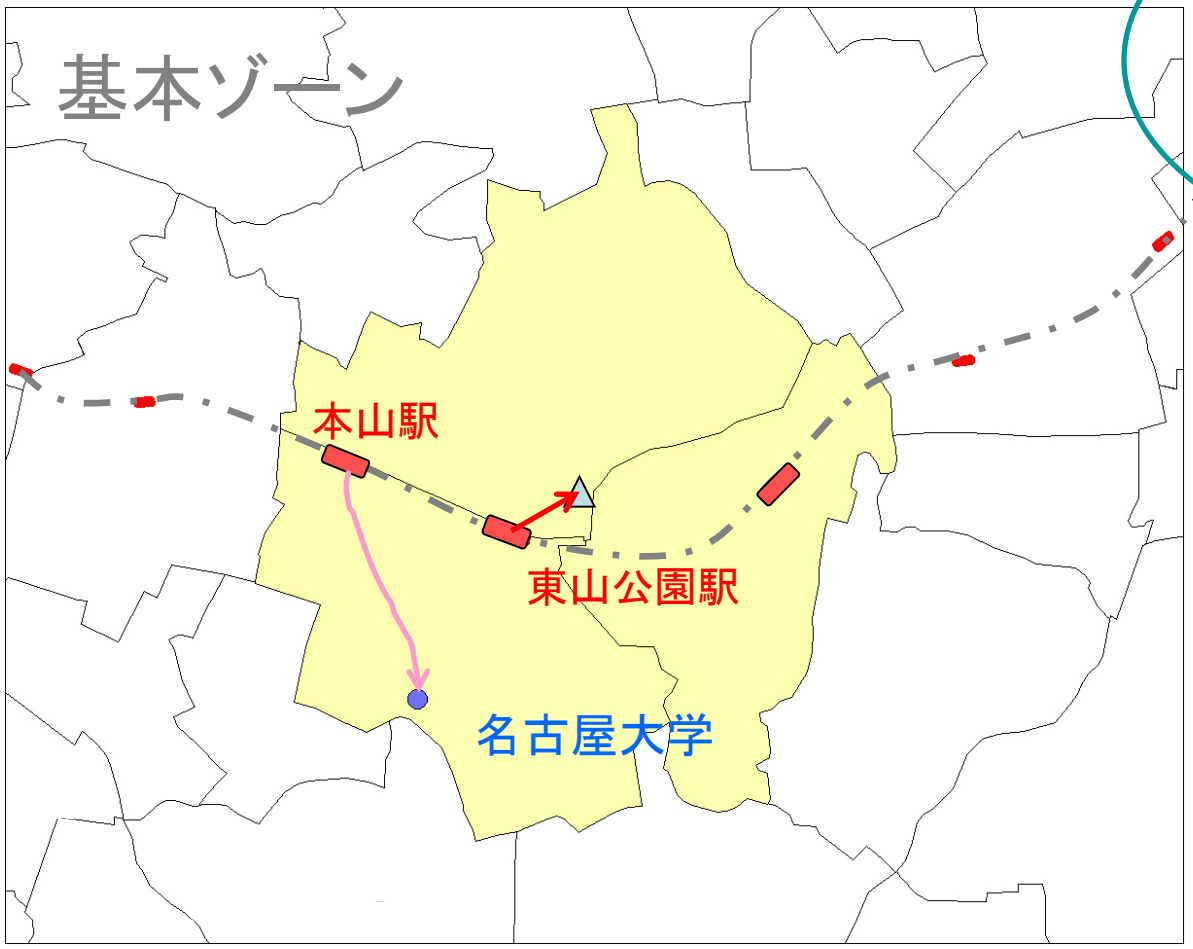
アクセス距離⇒潜在クラスモデルを導入
イグレス距離⇒施設を特定して実測値

分析対象

・公的施設（官公庁，病院，学校（大学））へのトリップ

・代表交通手段が鉄道，バス，自動車のトリップ

駅アクセス距離の算出



実測　⇒730m

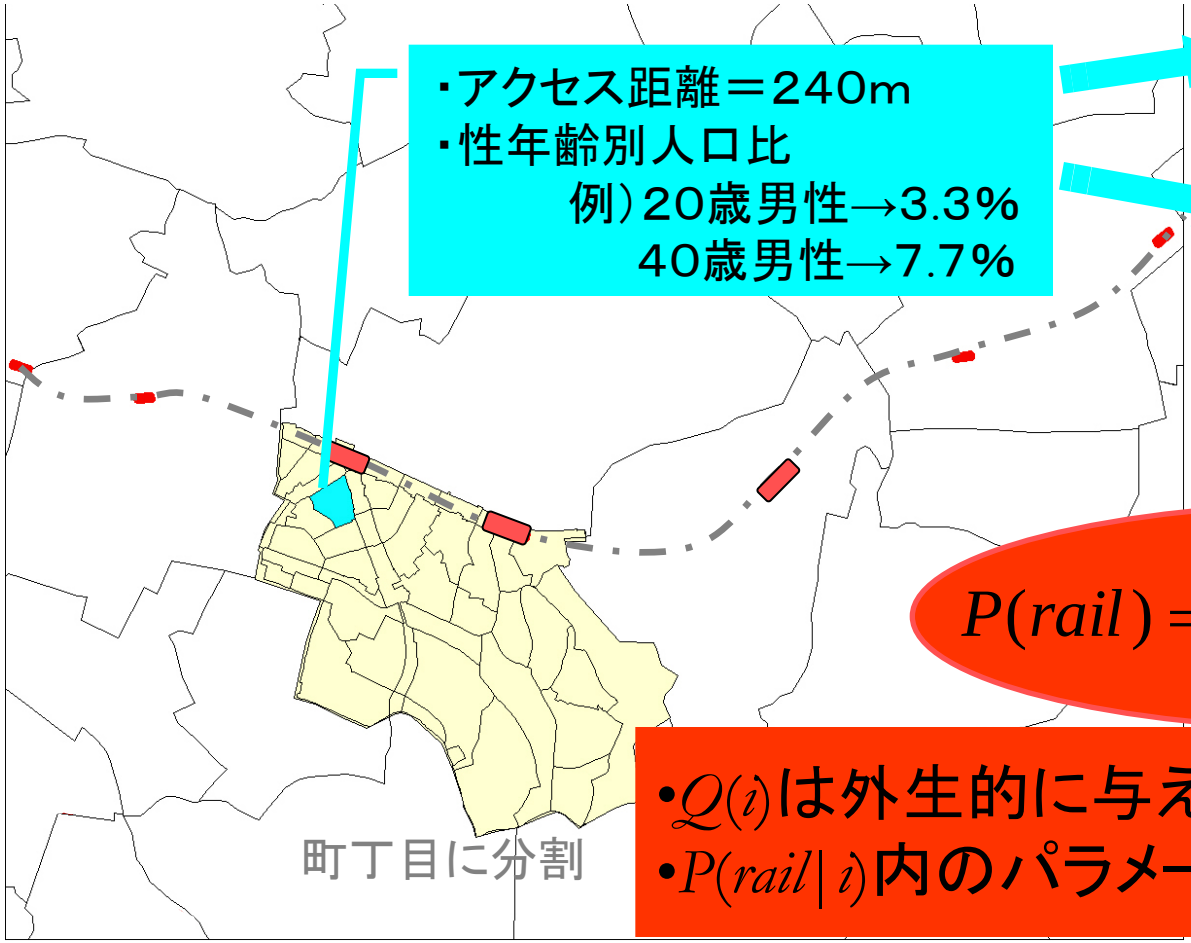
小ゾーン　⇒490m

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、



$P(rail \mid i)$ 鉄道の選択確率

$Q(i)$ 存在確率

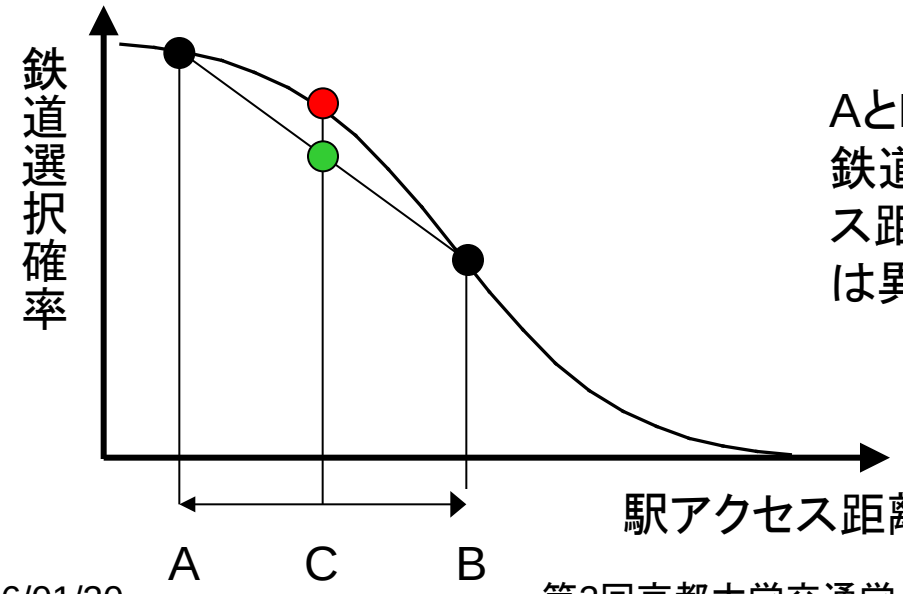
潜在クラスモデル

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

- $Q(i)$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

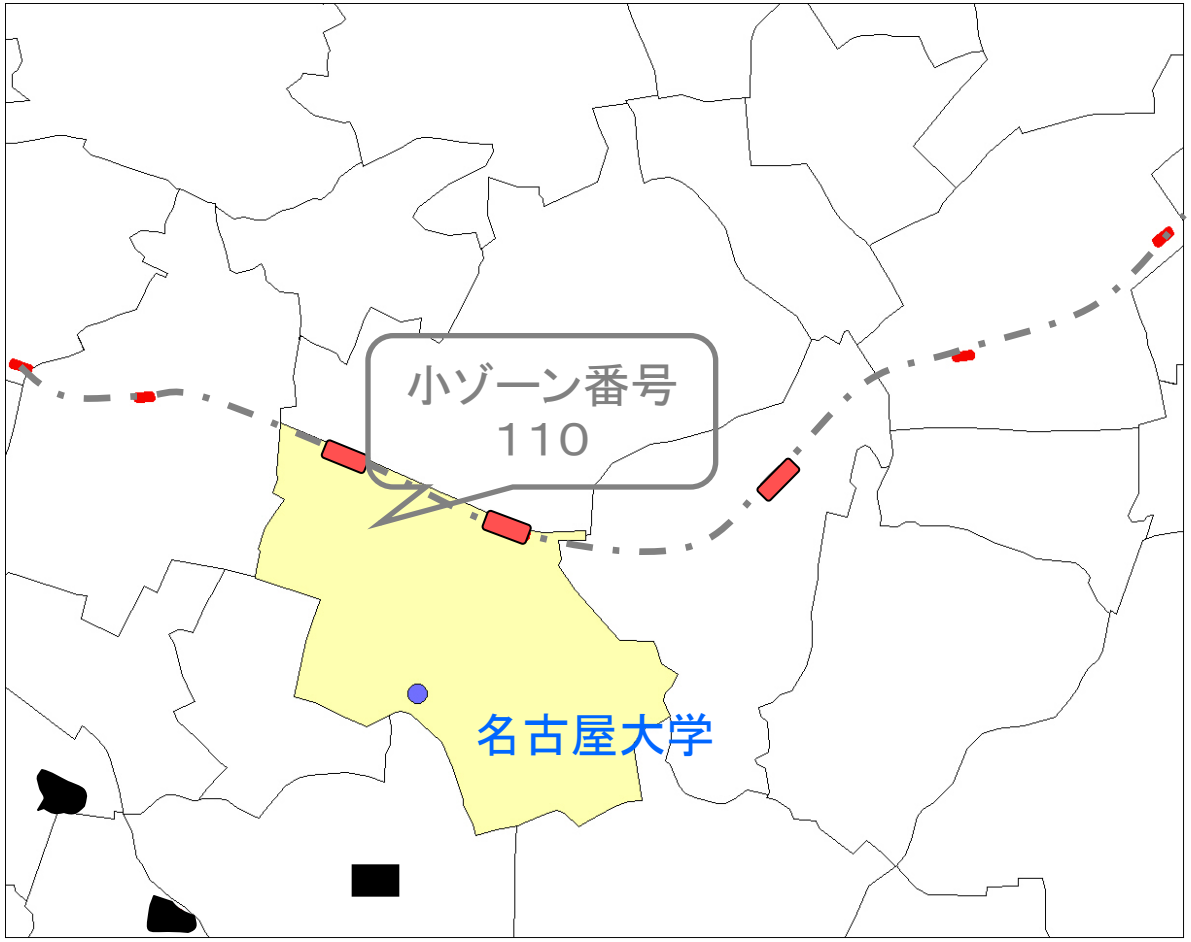
注意

- 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

イグレス距離の算出



目的地小ゾーン＝110
目的地施設　　＝学校
（18歳以上）

名古屋大学への
トリップと判断

同様のことを、目的地施設別にサンプル数の多い20個の小ゾーンについて行う

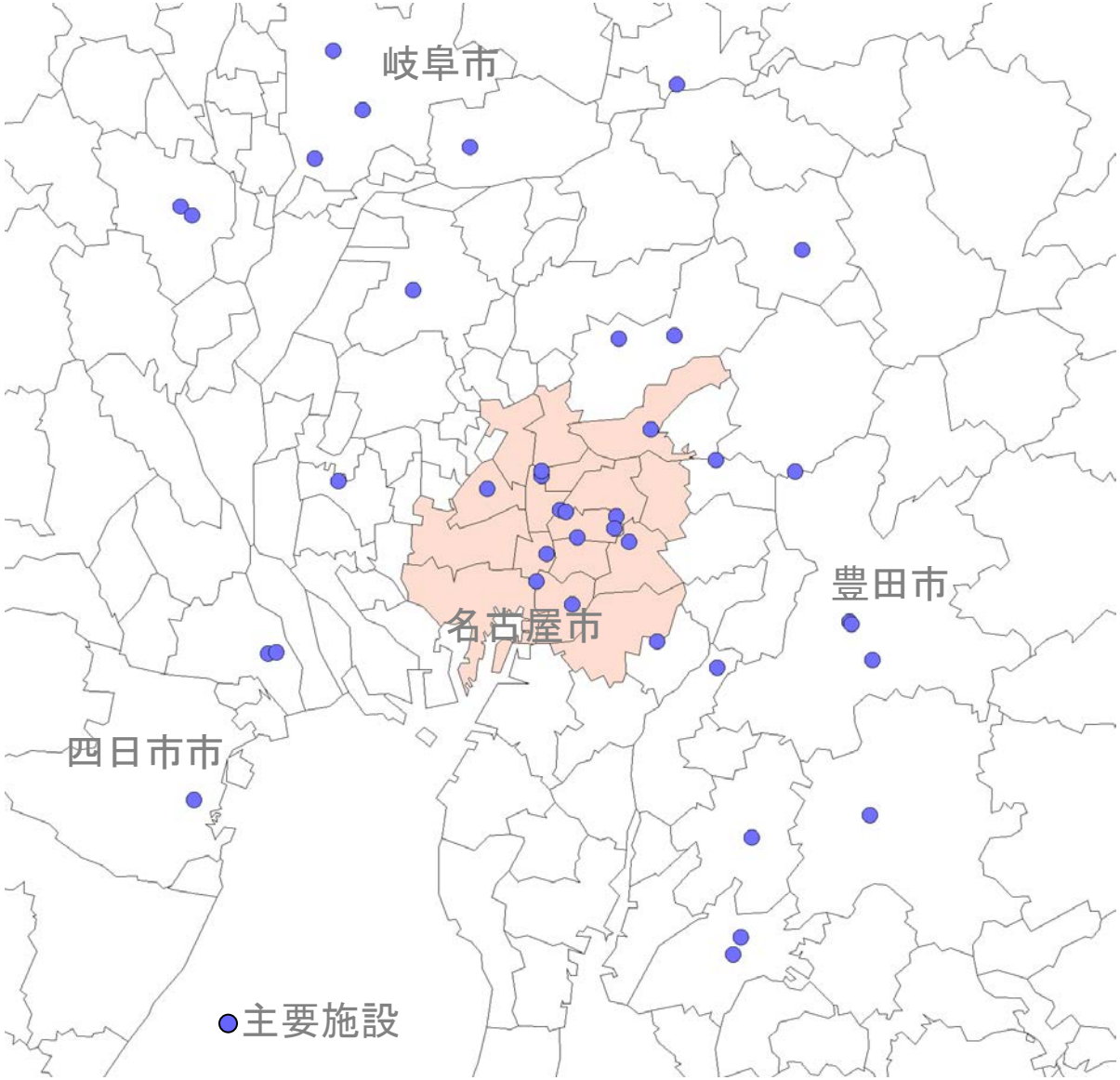
対象目的施設

PTゾーン範囲から

集中トリップ数の多い

60（20×3種類）を抽出

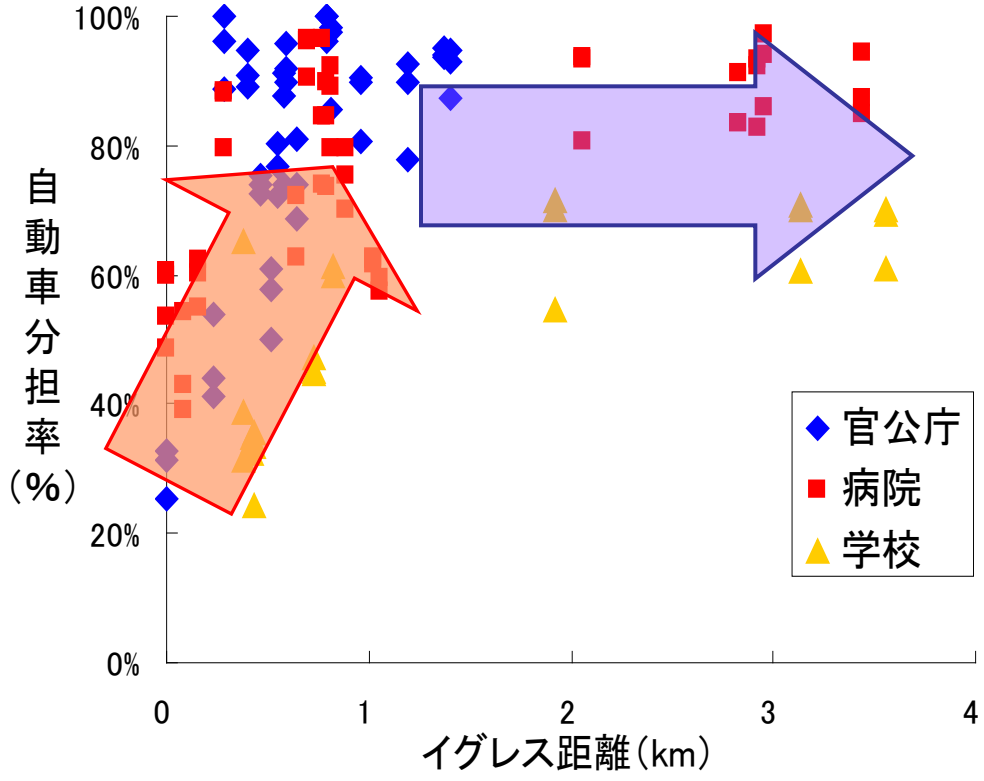
サンプルトリップ数4018



2006/01/20

集計分析①

イグレス距離と自動車分担率の関係



自動車分担率は

最寄り駅から離れるほど高くなる

1km以上の距離になると横ばい

集計分析②

公共施設のモータリゼーションへの対応

表-1　1970年以降に新規開設した施設の最寄り駅までの距離

	時期	距離(km)	順位(16中)
A病院	1984年	**4.9**	1位
B病院	1974年	**2.9**	4位
C病院	1972年	**2.8**	5位

表-2　1970年以降に移転した施設の最寄り駅までの距離の変化

		距離(km)	
	時期	移転前	移転後
D県庁	1966年	**1.1**	**1.4**
E市役所	1976年	**0.2**	**1.0**
F病院	1974年	**0.4**	**0.8**
G病院	1978年	**0.6**	**0.6**

最寄り駅から遠い

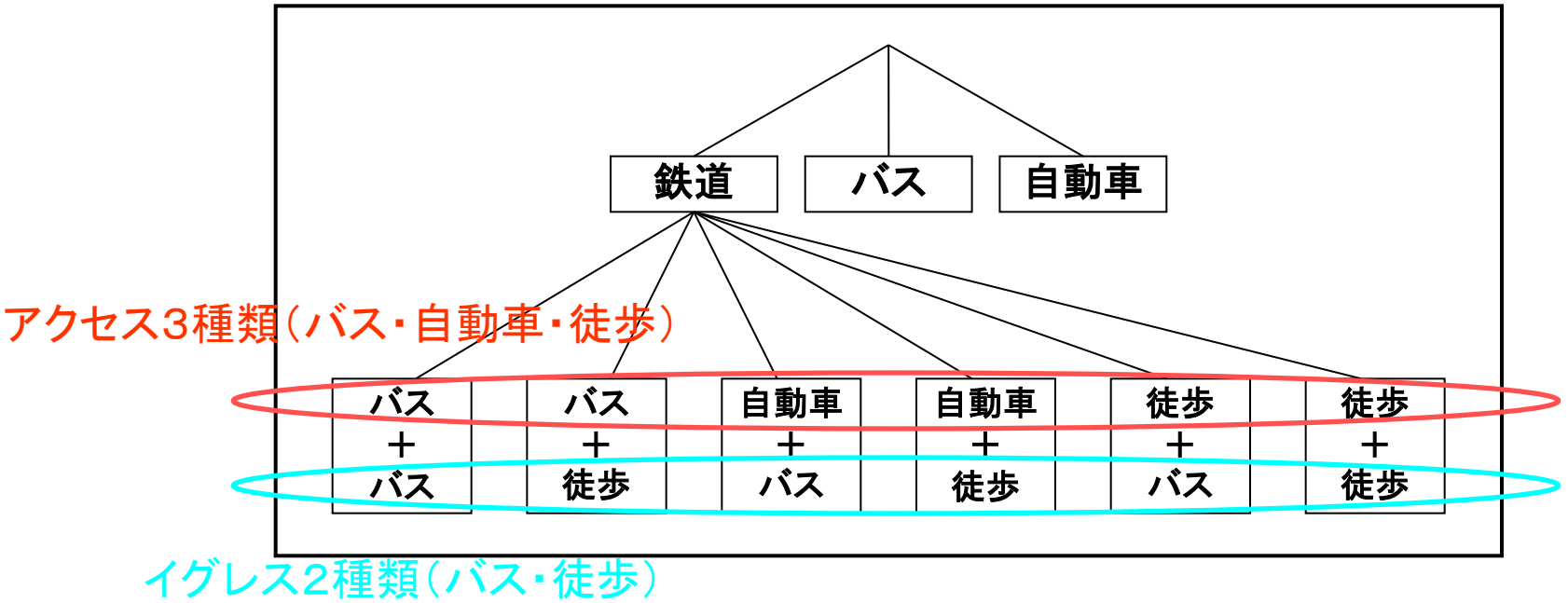
最寄り駅からの距離が増加

正のフィードバック

自動車交通の利用を促すような方向で推移

交通機関選択モデルの構築

ネスティットロジットモデル ➡ 鉄道の端末手段選択を下位段階に持つ



以下の組み合わせでモデルの有効性を検討

アクセス：潜在クラス，最小ゾーン，基本ゾーン
イグレス：実測値，　　最小ゾーン，基本ゾーン

イグレス観測精度の影響（アクセスは最小ゾーン）

	実測値	最小ゾーン
代表手段		
バス停イグレス	-2.0 (-3.4)	-1.7 (-4.2)
鉄道端末		
駅イグレス	-2.8 (-18.9)	-1.8 (-17.8)
最終尤度	-2944	-3063

実測値の適用により

- 適合度が向上
- パラメータ値が増大

アクセス観測精度の影響（イグレスは実測値）

	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

アクセスとイグレスが交通手段選択に及ぼす影響

	アクセス:潜在クラス イグレス:実測値	
代表手段		
バス停アクセス	-9.2	(-5.6)
バス停イグレス	-1.8	(-3.2)
鉄道端末		
駅アクセス	-1.4	(-11.8)
駅イグレス	-2.9	(-18.7)

- バス利用は自宅から最寄りバス停までの距離が支配的
- 駅端末選択ではイグレス距離の方がアクセス距離より影響が大きい

まとめ

- PTデータを用いてトリップ起終点位置観測精度の補完方法を提案，効果を統計的に確認
  - 起終点両方に潜在クラスモデルを用い，一般的なトリップへの適用可能性の検討が課題

- アクセスとイグレスで鉄道利用への影響は異なる
  - 自宅から駅までの遠さより駅から施設までの遠さの方が影響が大きい

ケース4：
観測の不完全性を考慮した自動車年間走行距離の分析

山本俊行，中川展孝：観測の不完全性を考慮した自動車年間走行距離の分析，土木計画学研究・講演集，Vol. 32, CD-ROM, 2005.

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

背景

自動車年間走行距離は

- 自動車利用を一定期間内で集計したもの
  - 自動車依存や世帯の交通パターンを表す重要な指標
  - ガソリン消費量や自動車の環境負荷，事故率等の算出にも用いられる

背景

しかし，通常，モデルの決定係数は低い

- 例: 0.11 (Train, 1986), 0.15 (Kockelman, 1997), 0.17 (Yamamoto et al., 2001)等

考えられる理由:

- 世帯間での自動車利用の異質性
  - 影響要因がモデルで十分考慮されていない
- 観測の不完全性
  - 被験者の記憶に頼る回顧データ
  - 1週間程度の交通行動履歴やオドメータによる観測

目的

- 観測の不完全性について明らかにする
- 観測の不完全性を考慮した年間走行距離モデルを構築する
  - モデルの有効性を検証する
- 被験者間の不完全性の異質性について明らかにする

データ

Parc-Auto

- フランスで実施されている世帯の自動車保有に関するパネル調査データ
- 1976年から毎年実施され現在も継続中
- サンプル数約7,000世帯，4年毎のローテーションパネル
- 3台までの保有自動車属性，自動車利用状況等を含む

年間走行距離の観測

2種類の観測

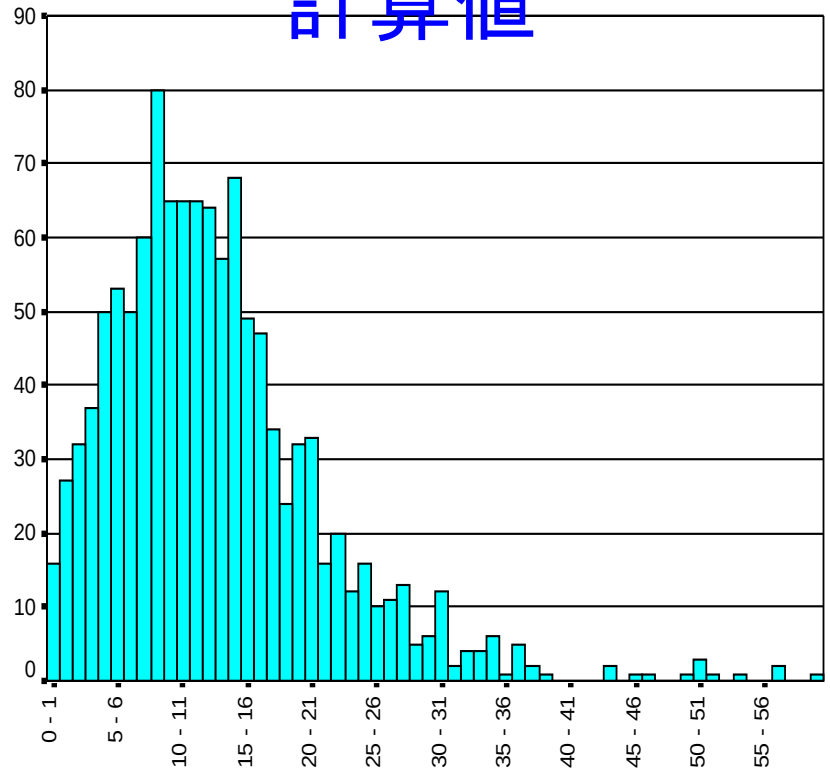
- 被験者の回答による年間走行距離:回答値
- 2年間のオドメータの差:計算値

以降では 両方の観測が得られた1167 ケースを用いた分析を行う

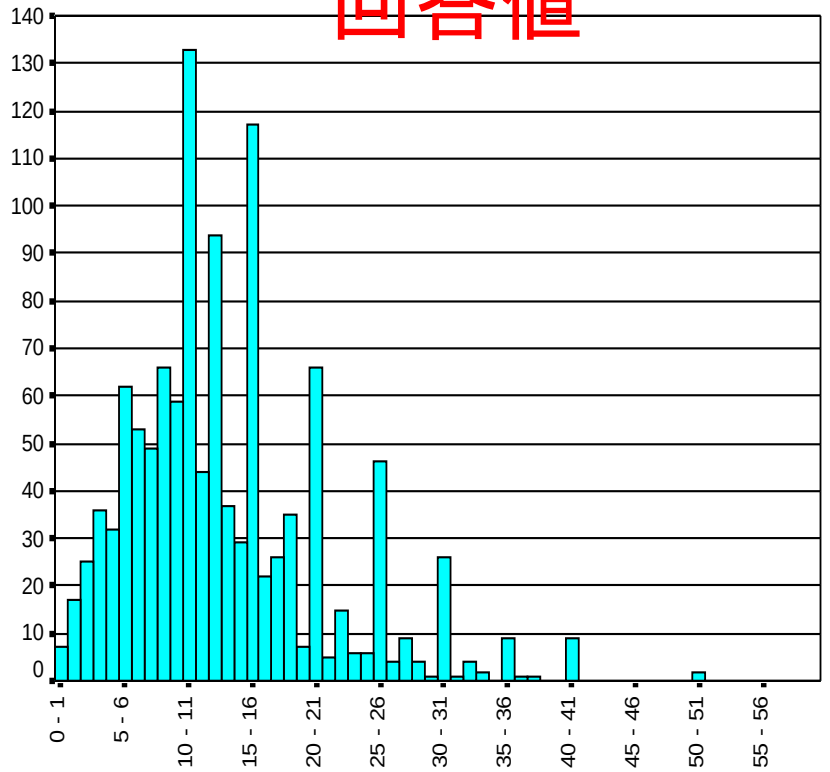
- 回答値:1998 年の回答値
- 計算値: 1997 年と1998 年のオドメータ

年間走行距離の分布

計算値

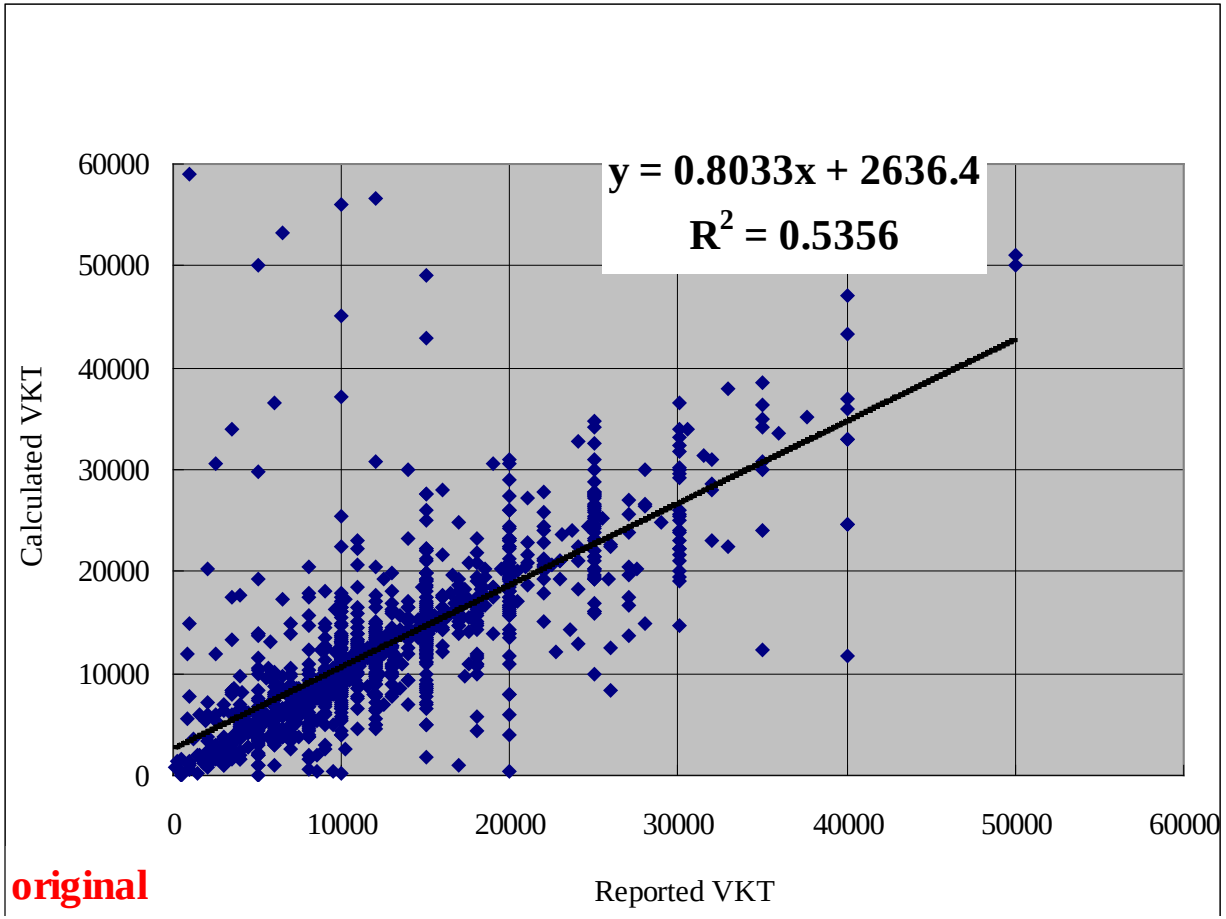


回答値



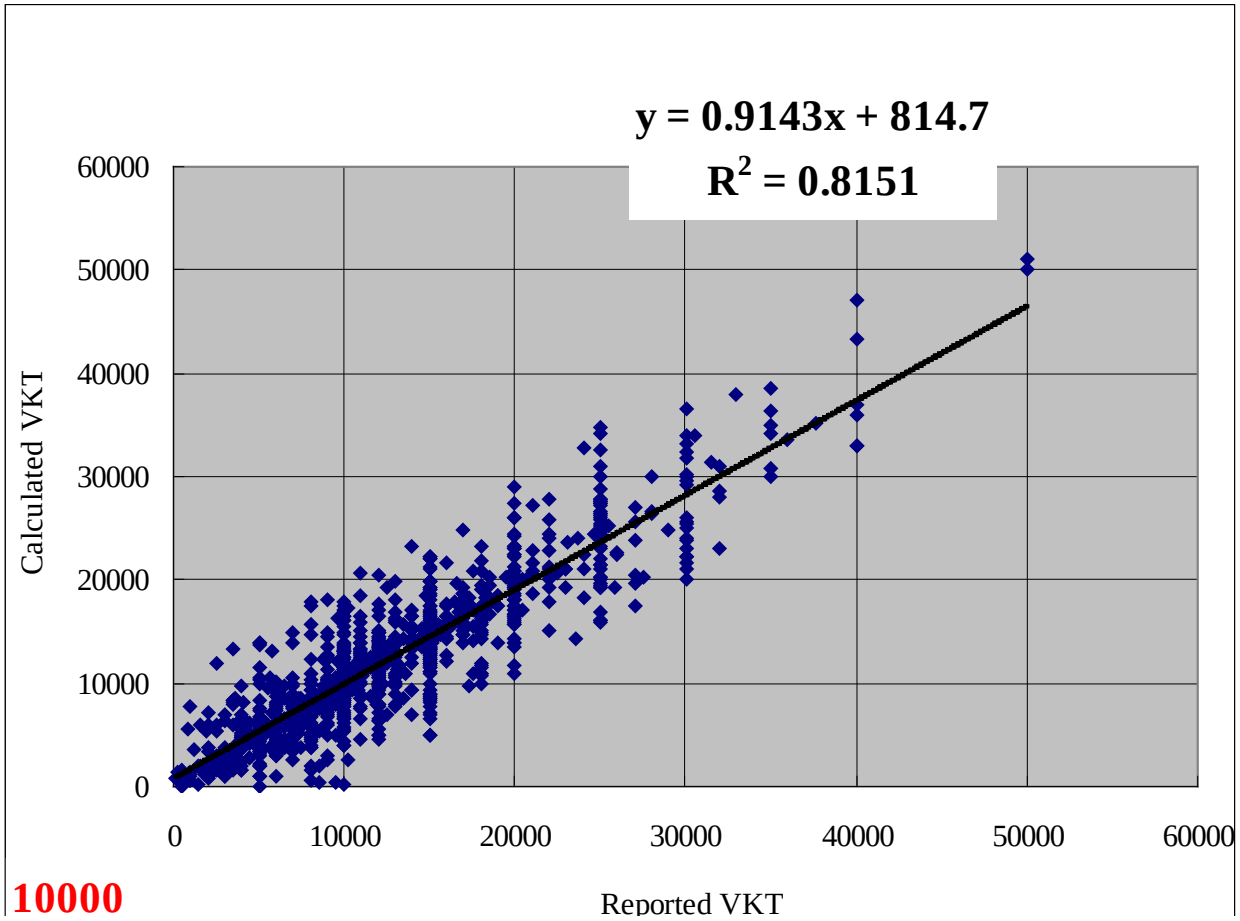
- 回答値には極端に台数が多い区間がある

計算値と回答値の分布



- いくつかのケースで乖離が非常に大きい

計算値と回答値の分布



- ところどころでプロットが縦にならんでいる

報告誤差の分析

- 計算値と回答値の差を被説明変数とした回帰分析
  - 差の絶対値
  - 回答値が計算値より小さい場合（過小報告）の差
  - 回答値が計算値より大きい場合（過大報告）の差

報告誤差の回帰モデル分析結果

- 保有台数が多い世帯は報告誤差が小さい
- 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す
- 年間走行距離が長いほど報告誤差が大きい

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モデル

$$y_i = \beta x_i + \varepsilon_i$$

- 個人間で一定の丸め誤差を仮定した場合

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

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

$$= \Phi\left(\frac{\mu_i - \boldsymbol{\beta}\mathbf{x}_i}{\sigma_\varepsilon}\right) - \Phi\left(\frac{\mu_{i-1} - \boldsymbol{\beta}\mathbf{x}_i}{\sigma_\varepsilon}\right)$$

個人間で不完全性には異質性がある

- 各個人の丸め誤差の大きさを表すモデル

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

$$z_i = 1 \quad \text{if } z_i^* < 0,$$

$$= 2 \quad \text{if } 0 \le z_i^* < \theta,$$

$$= 3 \quad \text{if } \theta \le z_i^*$$

  - 基礎分析の結果より500km単位，1000km単位，5000km単位の丸め誤差が混在していると仮定する
  - 年間走行距離そのものも丸め誤差に影響を及ぼす

⇒ 年間走行距離モデルと丸め誤差のモデルの誤差項に相関が生じる

誤差の相関を考慮した推定

- 年間走行距離モデル
- 丸め誤差モデル

$$y_i = \beta x_i + \varepsilon_i$$

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

- ε_i と ζ_i は無相関と仮定すると，z_i^* と y_i の誤差の共分散行列は

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

- 二変量プロビットモデルの枠組みで最尤推定を適用する

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

モデルの推定結果

年間走行距離モデル

- 通常の回帰モデルに比べて年間走行距離モデルの全体としての説明力が向上した
- 説明変数の推定値については通常の回帰モデルとそれほど変化しなかった

丸め誤差モデル

- 年間走行距離が長いほど丸め誤差が大きい
- 大型車は丸め誤差が大きい

まとめ

- 計算値と回答値の差の分析より
  - 世帯の保有台数や通勤への利用が回答誤差に影響する
  - 年間走行距離が長いほど丸め誤差が大きい
- 観測の不完全性を考慮した年間走行距離モデルの構築により
  - 通常の回帰モデルより説明力が向上した
  - 年間走行距離が長いほど丸め誤差が大きくなることを確認した
- 今後は
  - モデルによる分布の再現性についての検証
  - トリップ所要時間や出発時刻の分析への適用

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