The 23nd Behavior Modeling Summer School

Sep. 11 – 13 , 2024 @ The University of Tokyo

【日本語版】

Size Matters:

Or how to make your model useful for policy makers

Giancarlos Parady – The University of Tokyo

Parady, G., Ory, D., Walker, J. (2021) The overreliance on statistical goodness of fit and under-reliance on validation in discrete choice models: A review of validation practices in the transportation academic literature . Journal of Choice Modelling 38, 100257 (Open Access)

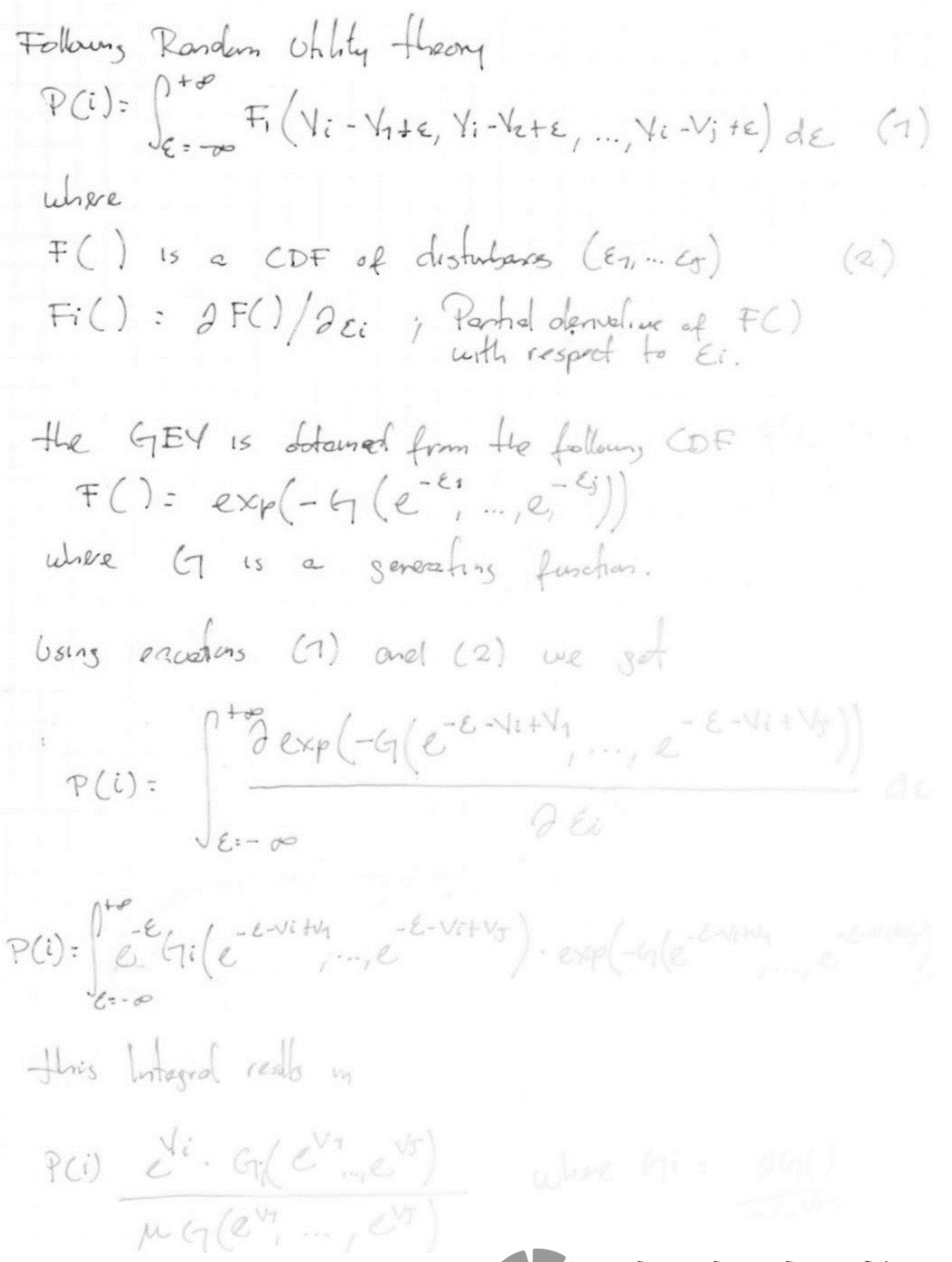
Parady, G., Axhausen K.W. (2023) Size Matters: The Use and Misuse of Statistical Significance in Discrete Choice Models in the Transportation Academic Literature. Transportation (accepted)

東京大学
THE UNIVERSITY OF TOKYO

Size Matters:

Or how to make your model useful for policy makers

Basic inference with discrete choice models



東京大学
THE UNIVERSITY OF TOKYO

Why is inference important?
具体的な事例：二項ロジットモデル

変数名		係数	標準誤差	t値
自動車定数項	ASC	1.45	0.393	3.70
乗車時間（分）	共通	-0.0089	0.0063	-1.42
乗車外時間（分）	共通	-0.0308	0.0106	-2.90
自動車out-of-pocket 費用 (c)	固有	-0.0115	0.0026	-4.39
公共交通運賃	固有	-0.0070	0.0038	-1.87
自動車保有ダミー（自動車）	固有	-0.770	0.213	3.16
市街地の勤務先ダミー（自動車）	固有	-0.561	0.306	-1.84
サンプル数		1476		
LL(0)		-1023		
LL(β)		-347.4		
-2[LL(0)-LL(β)]		1371		
ρ^2		0.660		
$\bar{\rho}^2$		0.654		

← 係数から効果の大きさを解釈できない

主に符号を解釈する

係数を用いて、効用や選択確率を求められる

効果の大きさを把握するため、弾力性か限界効果を求めるべき（重要！）

← 全ての係数が0の場合の尤度

← 最尤度

← Test of null hypothesis that all parameters are jointly zero. χ^2 distributed

← 適合度指標: 1 – (LL(β)/LL(0)

← 適合度指標: 1 – (LL(β)–K)/LL(0)

Source: Adapted from Ben-Akiva and Lerman (1984)

Basic Inference with discrete choice models

MNL・効果量の測定：点弾力性（Point elasticities）

Direct Elasticity

- **自己弾力性:** 選択肢 i の特性の**変化率（%）**に対する選択肢 i の選択確率の**変化率（%）**である。

  →選択肢 i に対する変数 x_{ink} が1%増えると、選択肢 i の選択率が何%変わるか。

$$E_{x_{ink}}^{P(i)} = \frac{\partial P_n(i)}{\partial x_{ink}} \cdot \frac{x_{ink}}{P_n(i)} = [1 - P_n(i)] x_{ink}\, \beta_k$$

Cross Elasticity

- **交差弾力性:**選択肢 j の特性の**変化率（%）**に対する選択肢 i の選択確率の**変化率（%）**である。

  →選択肢 j に対する変数 x_{jnk} が1%増えると、選択肢 i の選択率が何%変わるか。

$$E_{x_{jnk}}^{P(i)} = \frac{\partial P_n(i)}{\partial x_{jnk}} \cdot \frac{x_{jnk}}{P_n(i)} = -P_n(j) x_{jnk}\, \beta_k$$

Basic Inference with discrete choice models

MNL・効果量の測定：点弾力性（Point elasticities）

$x_{ink} = f^k(z_{ink})$ の場合

Direct Elasticity

- **自己弾力性:** 選択肢 i の特性の**変化率（%）**に対する選択肢 i の選択確率の**変化率（%）**である。

$$E_{x_{ink}}^{P(i)} = [1 - P_n(i)]\beta_k \cdot \frac{\partial f^k}{\partial z_{ink}} z_{ink}$$

従って、$x_{ink} = ln(z_{ink})$ の場合

$$E_{x_{ink}}^{P(i)} = [1 - P_n(i)]\beta_k \cdot \frac{\partial \ln(z_{ink})}{\partial z_{ink}} z_{ink} = [\mathbf{1} - \boldsymbol{P_n(i)}]\boldsymbol{\beta_k}$$

Basic Inference with discrete choice models

MNL・効果量の測定：点弾力性（Point elasticities）

- 前のスライドに見せた弾力性は個人の弾力性を示す
- 集計弾力性を求めるには、Probability Weighted Sample Enumeration法を用いる:

$$E_{x_{ink}}^{\overline{P(i)}} = \frac{\sum_{n=1}^{N} \hat{P}_{in}(i) E_{x_{ink}}^{P(i)}}{\sum_{n=1}^{N} \hat{P}_{in}(i)}$$

サンプル自己弾力性

$$E_{x_{jnk}}^{\overline{P(i)}} = \frac{\sum_{n=1}^{N} \hat{P}_{in}(i) E_{x_{jnk}}^{P(i)}}{\sum_{n=1}^{N} \hat{P}_{in}(i)}$$

サンプル交差弾力性

ここで、$\overline{\mathrm{P(i)}}$ は選択肢 i に対する集計確率であり，$\hat{\mathrm{P}}_{\mathrm{in}}(\mathrm{i})$ 個人nの選択肢 i に対する選択確率である

- 集計の場合は、交差弾力性が全ての選択肢 i に対して必ずしも同じではない
- ダミー変数の場合は、点弾力性の解釈に意味がないため普段求めない

Basic Inference with discrete choice models

NL・効果量の測定：点弾力性（Point elasticities）

$$P(j) = \frac{e^{V_j/\tau}}{e^{IV(i)}} \cdot \frac{e^{\tau IV(i)}}{\sum_{i=1}^{I} e^{\tau IV(i)}} \leftarrow \text{NL RUM2 specification}$$

Train(2009)

Direct Elasticity

- **自己弾力性:**

選択肢 j がネストに属しない場合

$$E_{x_{jnk}}^{P(j)} = [1 - P_n(j)]x_{jnk}\,\beta_k$$

選択肢 j がネストiに属する場合

$$E_{x_{jnk}}^{P(j)} = \left[\left(1 - P_n(j)\right) + \left(\frac{1}{\tau} - 1\right)\left(1 - P(j|i)\right)\right] x_{jnk}\,\beta_k$$

Cross Elasticity

- **交差弾力性:**

選択肢 j とj' が違うネスト属する場合

$$E_{x_{j'nk}}^{P(j)} = -P_n(j')x_{j'nk}\,\beta_k$$

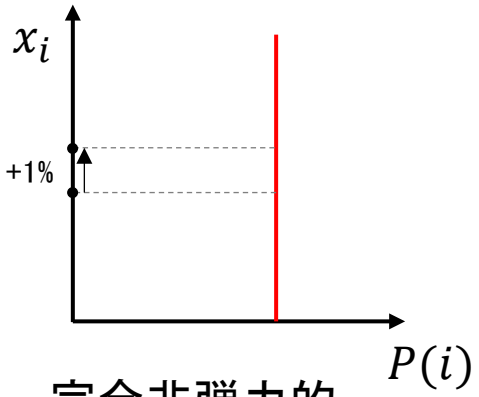
選択肢 j とj' が同じネスト属する場合

$$E_{x_{j'nk}}^{P(j)} = -\left[P_n(j') + \left(\frac{1}{\tau} - 1\right)P(j'|i)\right] x_{j'nk}\,\beta_k$$

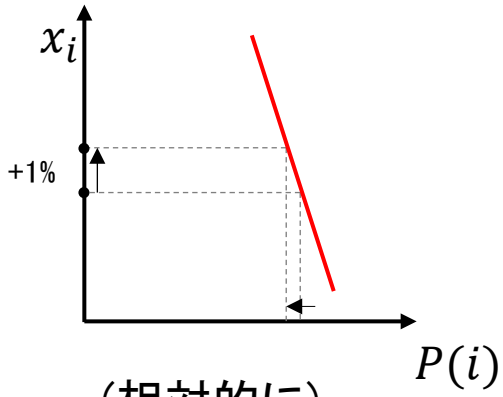
Basic Inference with discrete choice models

自己弾力性

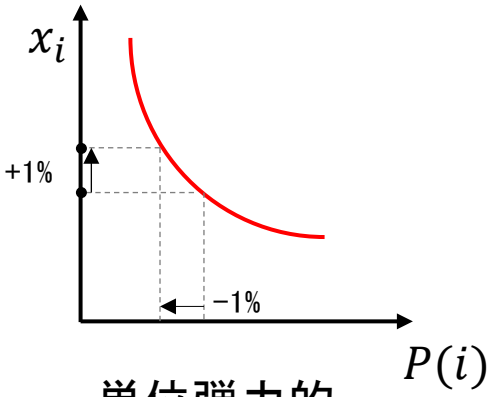
x_i は選択肢 iの費用とする



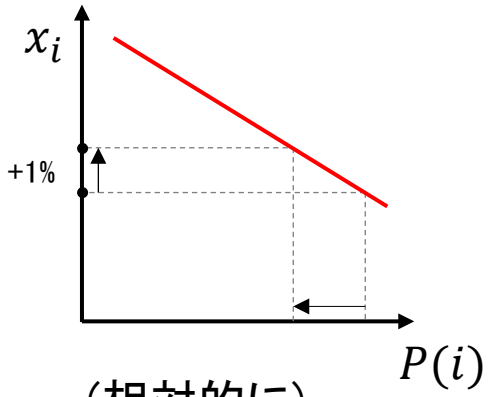
完全非弾力的



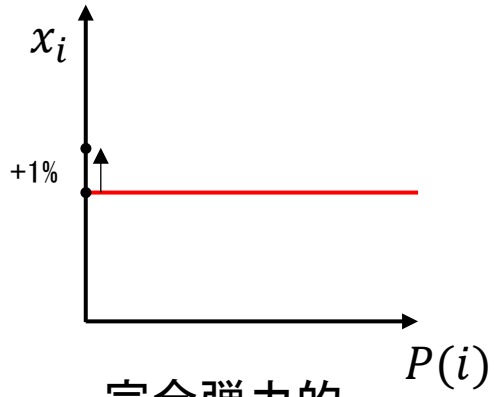
(相対的に)
非弾力的



単位弾力的



(相対的に)
弾力的



完全弾力的

自己弾力性:

x_iが1%増加したとき、
$P(i)$が0%減少する

x_iが1%増加したとき、
$P(i)$が1%未満減少する

x_iが1%増加したとき、
$P(i)$が丁度1%減少する

x_iが1%増加したとき、
$P(i)$が1%以上減少する

x_iが1%増加したとき、
$P(i)$が無限に減少する

交差弾力性:

x_jが1%増加したとき、
$P(i)$が0%増加する

x_jが1%増加したとき、
$P(i)$が1%未満増加する

x_jが1%増加したとき、
$P(i)$が丁度1%増加する

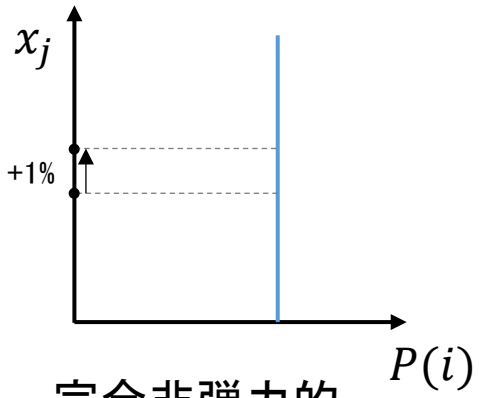
x_jが1%増加したとき、
$P(i)$が1%以上増加する

x_jが1%増加したとき、
$P(i)$が無限に増加する

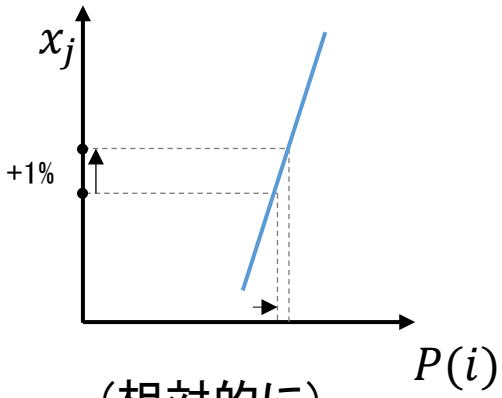
Basic Inference with discrete choice models

交差弾力性

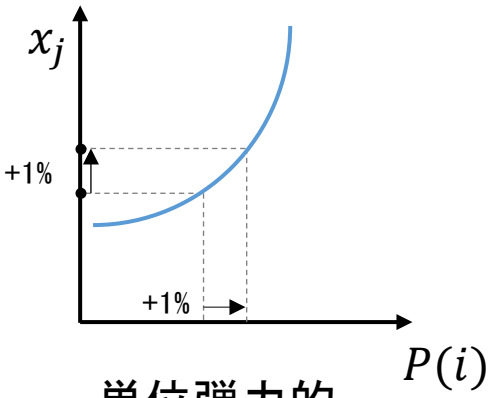
x_j は選択肢 j の費用とする



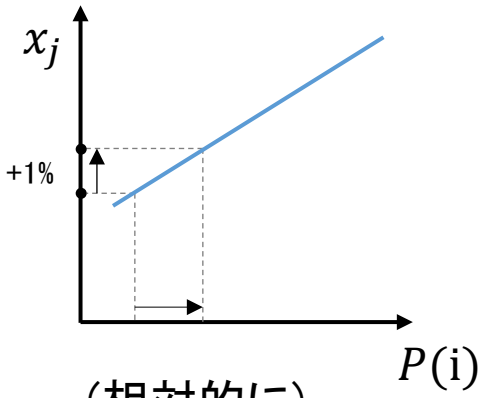
完全非弾力的



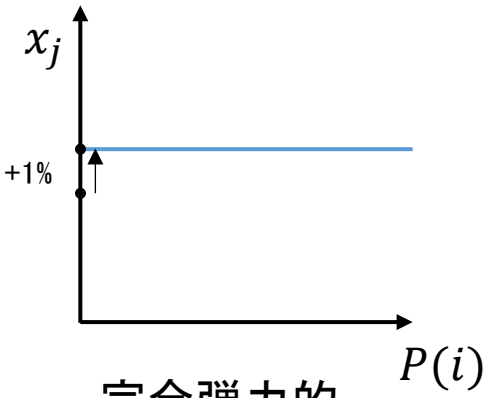
(相対的に)
非弾力的



単位弾力的



(相対的に)
弾力的



完全弾力的

自己弾力性:

x_iが1%増加したとき、$P(i)$が0%減少する	x_iが1%増加したとき、$P(i)$が1%未満減少する	x_iが1%増加したとき、$P(i)$が丁度1%減少する	x_iが1%増加したとき、$P(i)$が1%以上減少する	x_iが1%増加したとき、$P(i)$が無限に減少する

交差弾力性:

x_jが1%増加したとき、$P(i)$が0%増加する	x_jが1%増加したとき、$P(i)$が1%未満増加する	x_jが1%増加したとき、$P(i)$が丁度1%増加する	x_jが1%増加したとき、$P(i)$が1%以上増加する	x_jが1%増加したとき、$P(i)$が無限に増加する

Basic Inference with discrete choice models

MNL・効果量の測定：限界効果（Marginal effects）

Direct Marginal Effect

- **自己限界効果:** 選択肢 i の特性の**単位変化**に対する選択肢 i の選択確率の**変化量**である。

  →選択肢 i に対する変数 x_{ink} が1単位増えると、選択肢 i の選択率が何百分率点変わるか。

$$M_{x_{ink}}^{P(i)} = \frac{\partial P_n(i)}{\partial x_{ink}} = P_n(i)[1 - P_n(i)]\beta_k$$

Cross Marginal Effect

- **交差限界効果:**選択肢 j の特性の**単位変化**に対する選択肢 i の選択確率の**変化量**である。

  →選択肢 i に対する変数 x_{jnk} が1単位増えると、選択肢 i の選択率が何百分率点変わるか。

$$M_{x_{jnk}}^{P(i)} = \frac{\partial P_n(i)}{\partial x_{jnk}} = P_n(i)(-P_n(j)\beta_k)$$

Definition following Louviere, Hensher, and Swait (2000)

Basic Inference with discrete choice models

MNL・効果量の測定：限界効果（Marginal effects）

- 弾力性と同様にProbability Weighted Sample Enumerationを用いて集計限界効果を求められる:

$$M_{x_{ink}}^{\overline{P(i)}} = \frac{\sum_{n=1}^{N} \hat{P}_{in}(i) M_{x_{ink}}^{P(i)}}{\sum_{n=1}^{N} \hat{P}_{in}(i)}$$

サンプルの自己限界効果

$$M_{x_{jnk}}^{\overline{P(i)}} = \frac{\sum_{n=1}^{N} \hat{P}_{in}(i) M_{x_{jnk}}^{P(i)}}{\sum_{n=1}^{N} \hat{P}_{in}(i)}$$

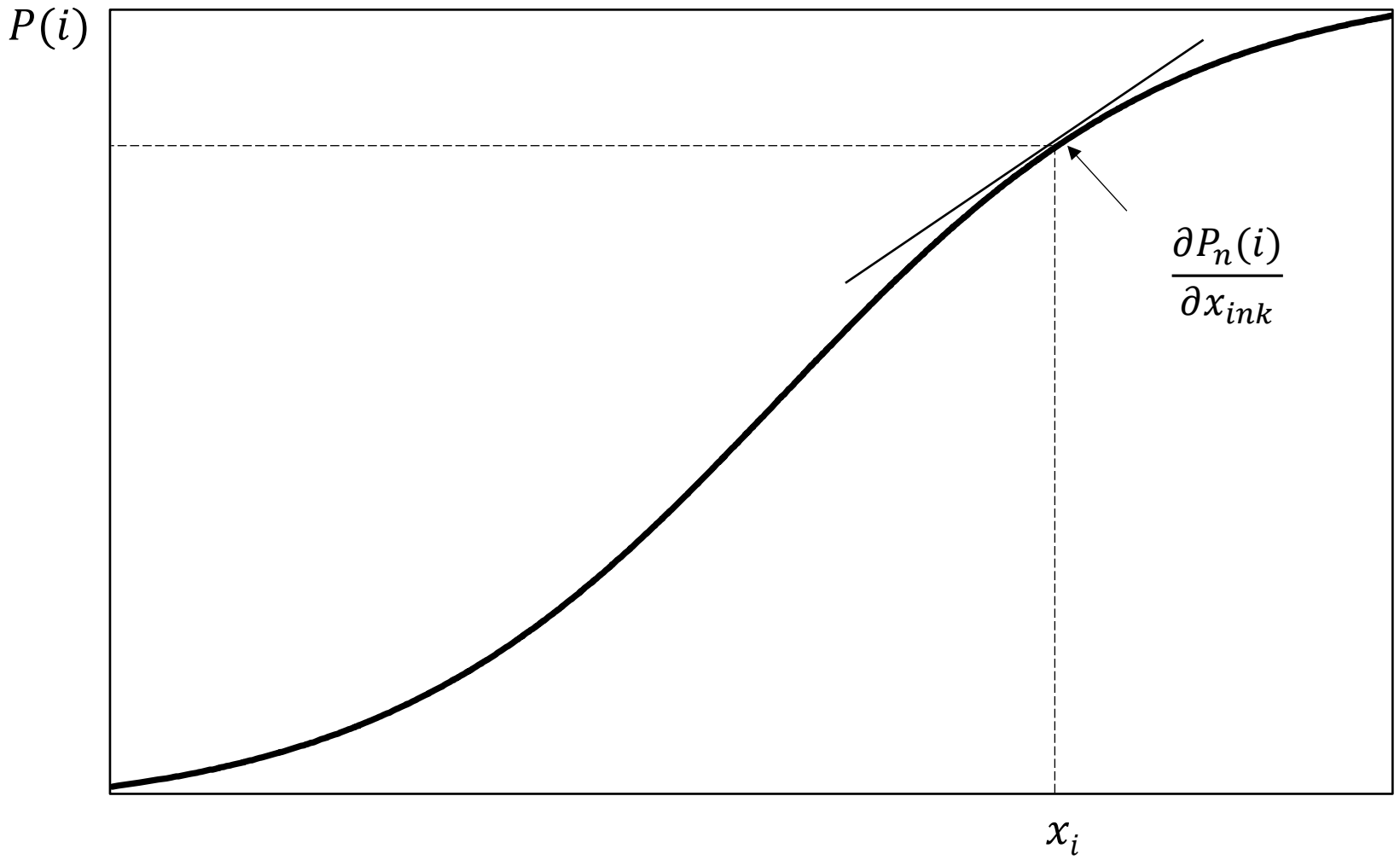
サンプルの交差限界効果

ここで、$\overline{\mathrm{P(i)}}$ は選択肢 i に対する集計確率であり，　$\hat{\mathrm{P}}_{\mathrm{in}}(\mathrm{i})$ 個人nの選択肢 i に対する選択確率である

- 限界効果の場合は、ダミー変数でも解釈に意味があるが求め方は異なる。

Basic Inference with discrete choice models

MNL・効果量の測定：限界効果（Marginal effects）



Marginal effects as the slopes of the Tangent lines to the cumulative probability curve

Adapted from Hensher, David A., John M. Rose, and William H. Greene. *Applied choice analysis: a primer*. Cambridge University Press, 2015 (2nd Edition)

Basic Inference with discrete choice models

効果量の測定：限界効果（Marginal effects）

■シミュレーションを用いて**ダミー変数における集計限界効果の求め方**

対象変数を0とおいて、ベース予測値を求める（個人レベル）

↓

対象変数を1とおいて、新予測値を求める（個人レベル）

↓

新予測値とベース予測値の差の平均値を求める

Basic Inference with discrete choice models

効果量の重要性は当たり前だろう？

Size Matters:

Or how to make your model useful for policy makers

ただし。。。

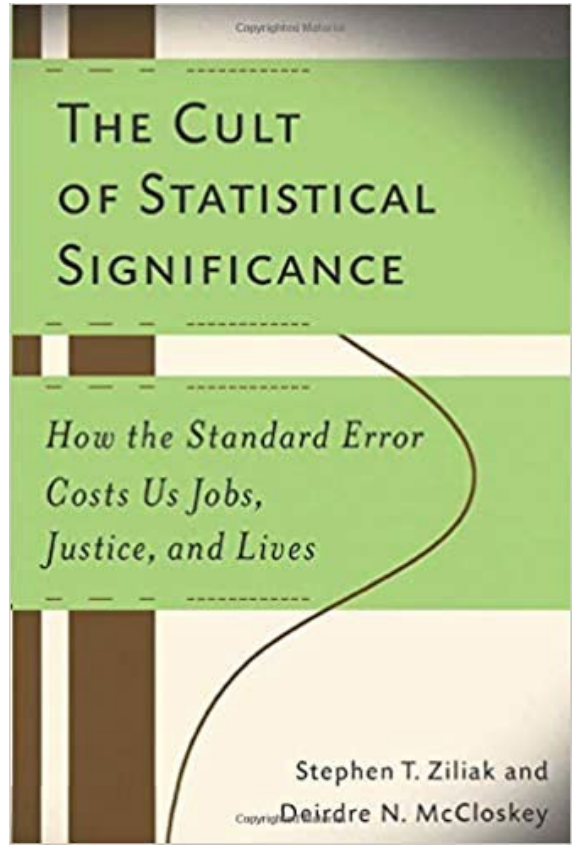
定量的研究を行う研究者は統計的有意性にこだわり過ぎる

東京大学
THE UNIVERSITY OF TOKYO

The Use and Misuse of Statistical Significance in Discrete Choice Models

A review of reporting practices in the Transportation Academic Literature

- 統計モデルは，**安価な計算力により，交通関連現象を説明するために重要な道具**となっている
- 定量的分析に頼る多くの分野において，統計モデルの普及とともに，（**政策変数の**）**実質的な重要性を考慮せず統計的有意性のみによる評価**も広まった
- McCloskey & Ziliakは，経済学の分野において，**統計的有意性の度重なる誤用を明らかにした**. しかし，これは**決して経済学に限らない**



本研究では，McCloskey and Ziliakの19の質問を離散選択モデル中心に交通計画分野の学術文献に適用し，**実証分析における統計的有意性の使用と誤用を評価する**.

交通計画分野における15質問

研究は…

Q1.モデルに使う変数の記述統計や単位を報告したか

Q2.「効果がどの程度大きいか」の問に応える弾力性，限界効果，または他の関心のある指標を報告したか

Q3. 標準誤差，t−値及び尤度比をすべて報告したか

Q4.検定の検出力が考慮したか

Q5.その場合は，検出力をどう扱ったか

Q6.「Asterisk econometrics」と呼ばれる検定統計量の絶対値の大きさによって係数をランク付けすることが回避したか

Q7. モデル結果の節で，「sign econometrics」と呼ばれる効果量を考慮せず係数の符号を解釈することを回避したか．

Q8. 効果量について議論したか．

交通計画分野における15質問

研究は…

Q9.効果量に対して，実質的な重要性について判断したか．つまり，実質的に効果のある要因とそうでもない要因を指摘しているか

Q10.効果量に対する実質的な重要性を判断するための科学的な根拠について議論したか

Q11. 統計的有意性のみに基づく変数の選択を回避したか

Q12.効果の妥当性の判断，または効果量をより良く説明するために，シミュレーションを行ったか

Q13.結論と政策への示唆の節で，統計的有意性と経済的，政策的及び科学的な重要性を区別したか

Q14. 推定，結論，政策への示唆の節では，あるときは「統計的」な意味で，別のときは「政策や科学で重要になるほど大きい」という意味で，「有意」という言葉を曖昧に使うことを回避したか

Q15.効果量の信頼区間を報告した且つ，統計的有意性の点推定値の代わりではなく実質的な重要性を解釈するために使用したか

The Use and Misuse of Statistical Significance in Discrete Choice Models

A review of reporting practices in the Transportation Academic Literature

主な調査結果（括弧内の数値はMcCloskey & Ziliak[2)] が報告した数値を示す）：

- **67%** (MZ:70%)の論文は，**統計的有意性と経済的，政策的，科学的な重要性を区別していない**
- **86%** (MZ: 72%)の論文は，**ある効果量が「大きい」か「小さい」かと判断できるように科学的な根拠について議論していない**
- **62%** (MZ: 59%)の論文は，「**有意」という言葉を**，あるときは帰無仮説と統計的に異なるという意味で，またあるときは実質的に重要であるという意味で，**曖昧に使っている**
- **39%** (MZ: 53%) の論文は，**係数の符号のみに基づいてモデル結果を説明している**（Sign Econometrics）
- **24%** (MZ: 32%) の論文は，**モデルから変数を除外するための唯一の基準として，統計的有意性を利用したことを明示している**
- 検定の統計的検出力を考慮した論文はなかった（MZ: 4%）

The Use and Misuse of Statistical Significance in Discrete Choice Models

A review of reporting practices in the Transportation Academic Literature

統計的有意性と係数の符号だけを踏まるモデル結果の議論

統計的有意性はただ一種の誤差を評価することであり、変数の**実質的な重要性の評価ではない！**

サンプルが十分大きいければ**なんでも有意になる**

“変数 X は有意で正な影響を与えている…”

正か負はどうでもいい！効果量は考慮しない限り意味のない情報だる

サイズが重要だ！

小さなはワイングラスは
だれも欲しくない！

The Use and Misuse of Statistical Significance in Discrete Choice Models

A review of reporting practices in the Transportation Academic Literature

（有意性ではなく）効果量に中心する議論＋実質的な重要性についての判断

Khan, Kockelman and Xiong (*2014*)が，**「(0.5マイル以内の十字路として定義した)ネットワークの接続性が大きな影響を与える:この変数の標準偏差1つの増加は，徒歩の選択率を34%増加させると推定される」**と述べて，さらに**「駐車料金や無料駐車場の有無の変数は，あまり効果がないことが分かった」**と述べており，効果量を明確に判断している.

de Luca and Di Pace (2015)は，交通時間価値の推定値の議論の中で駐車場立地の実質的な重要性の判断を明確にしており，**「イタリアにおける別の研究で報告された推定値と同程度であり，駐車場の立地が極めて重要であることを示している. 片道の平均交通費を3ユーロと仮定すると，10分間の歩行時間(時速4kmで約700m)は，総交通費の半分以上となる.」**と述べている.

The Use and Misuse of Statistical Significance in Discrete Choice Models

A review of reporting practices in the Transportation Academic Literature

統計的有意性と実質的な重要性の混同

意思決定過程における社会的相互作用の影響に関する研究で，Kamargianni et al. (2014)は，歩行嗜好の潜在的構成要素について，**「この構成要素が最も統計的に有意な変数であり…親が子どもの歩行に対する態度の発達に強い影響を与えることを示している」**と述べており，大きなt-統計量を大きいな効果量と誤解している.

Qin et al. (2017)は交通手段転換に関する研究で，**「バスのサービスレベルが最も有意な正のt-値を有しており，バスのサービス水準を向上させることで，自動車利用者に対するバスへの転換率を有意に増加させることができることを示している」**と論じている.

The Use and Misuse of Statistical Significance in Discrete Choice Models

A review of reporting practices in the Transportation Academic Literature

改善策の提案

①効果量とその信頼区間の報告を義務付ける

- **統計的有意性は様々な評価基準の中で，たかが一つであり，最も重要な基準であってはならない**．そこで，統計モデルの議論は，効果量または他の政策に関連する指標に焦点を当てるべきである．
- モデル係数を報告することには反対はしないが，直接的に解釈できず，論文の付録とすることで十分であろう．

②可能な限り，効果の大きさについて判断したうえで（著者によるどの効果が「小さい」かどの効果が「大きい」か），その判断の根拠を報告する．

- どの程度の効果が政策に対して重要になるか，その重要性をどう評価するかの議論と共に，政策変数の操作をかかるコスト（政策導入コスト）を議論すべきである．

The Use and Misuse of Statistical Significance in Discrete Choice Models

A review of reporting practices in the Transportation Academic Literature

改善策の提案

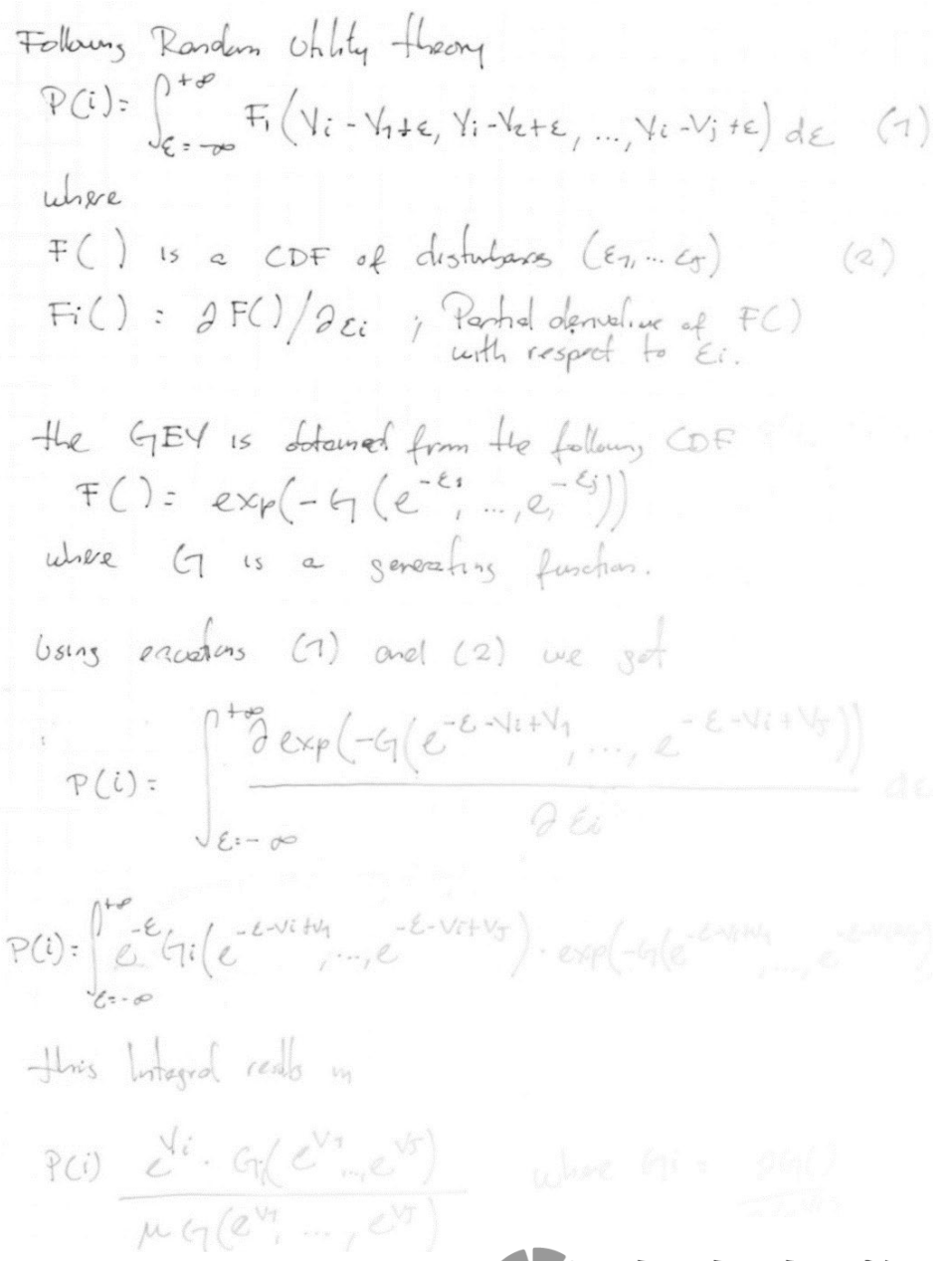
③可能な限り，推定した効果量や関心のある指標を既往研究と比較する.

- 交通時間価値等よく報告されている値の場合は，既往研究との比較への障壁はない.
- よく報告されていない値の場合，変数の定義及び測り方のばらつきにより，比較することが難しくなる時もある．ただし，効果量の報告が普及されたらこの問題が解消される.

④新規研究については，対象効果を十分な検出力で検出できることを保証するために，サンプルサイズを決める際に統計的検出力を考慮すること．二次データ（例：パーソントリップ調査など）を用いた研究の場合は，事後に各検定の検出力を求めて報告する.

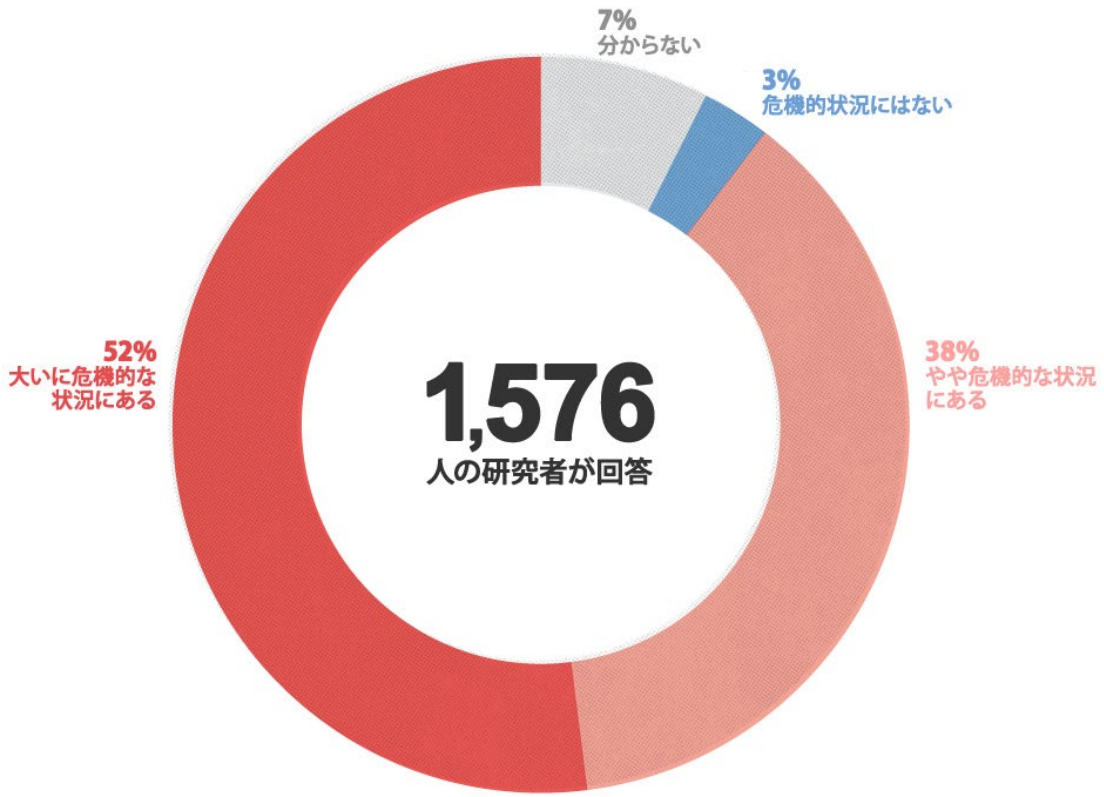
自習用:

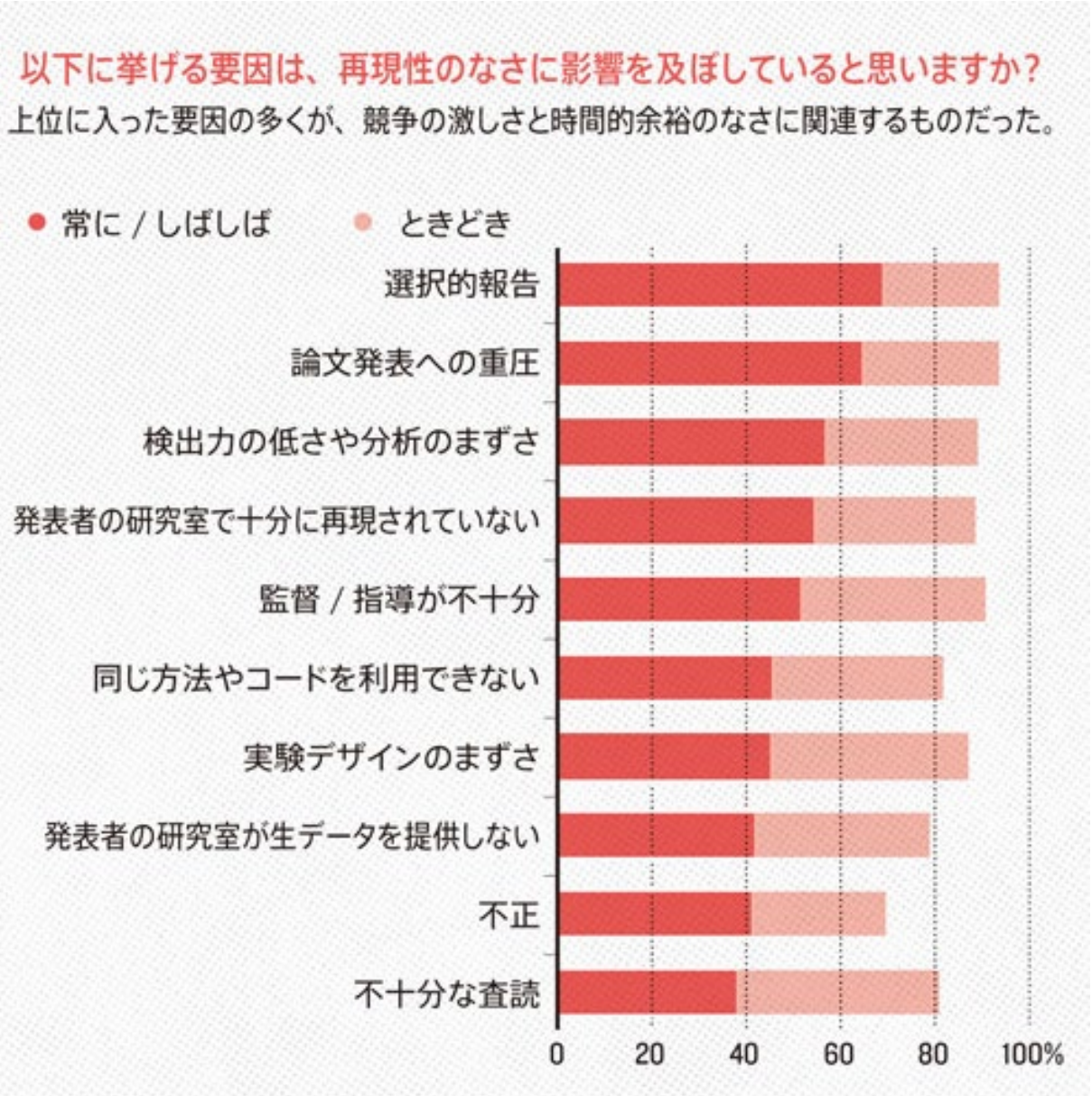
Validation practices in discrete choice modeling



東京大学
THE UNIVERSITY OF TOKYO

A credibility crisis in science and engineering?





Source: Baker, M. and Penny, D. (2016) 'Is there a reproducibility crisis?', *Nature*, 533 (7604)

A credibility crisis in science and engineering?

統計的検出力のなさ、小さい効果量及び、研究実験設計、定義、成果と方法論の緩みによって、発表された研究成果のほとんどは偽である。

(Ioannidis, 2005)

Focused on experimental studies

自然科学と異なり、

- 横断的な観察研究に依存する。
- 仮説を科学的に検定することは難しい（反証可能性の検証）。
- 以上は丁寧なモデル検証の必要性を強調する。

→実践において、予測誤差は測定可能であるため、予測に対するフィードバックを受けられるが、アカデミアにおいてはそのフィードバックがほとんどない。

主要な用語定義

予測精度（Predictive accuracy）:予測と実際観測されたアウトカムの一致度合い。

予測精度の要素：

- **較正（Calibration）:** 予測された確率と観測されたアウトカムの相対頻度の一致度合い。
- **判別能力（Discrimination ability）:** モデルによる各個体に対するアウトカムの有無の判別能力。

汎化性能（Generalizability）: 未知データ（推定に用いていないデータ）において予測精度の維持能力。

汎化性能の要素：

- **再現性（Reproducibility）:** 同じ母集団から抽出された別の標本において予測精度の維持度合い。
- **移転性（Transferability）:** 尤もらしく類似性のある別の母集団から抽出された標本または、異なる方法で収集されたサンプルにおいて予測精度の維持度合い。

Parady, Ory & Walker (2021)

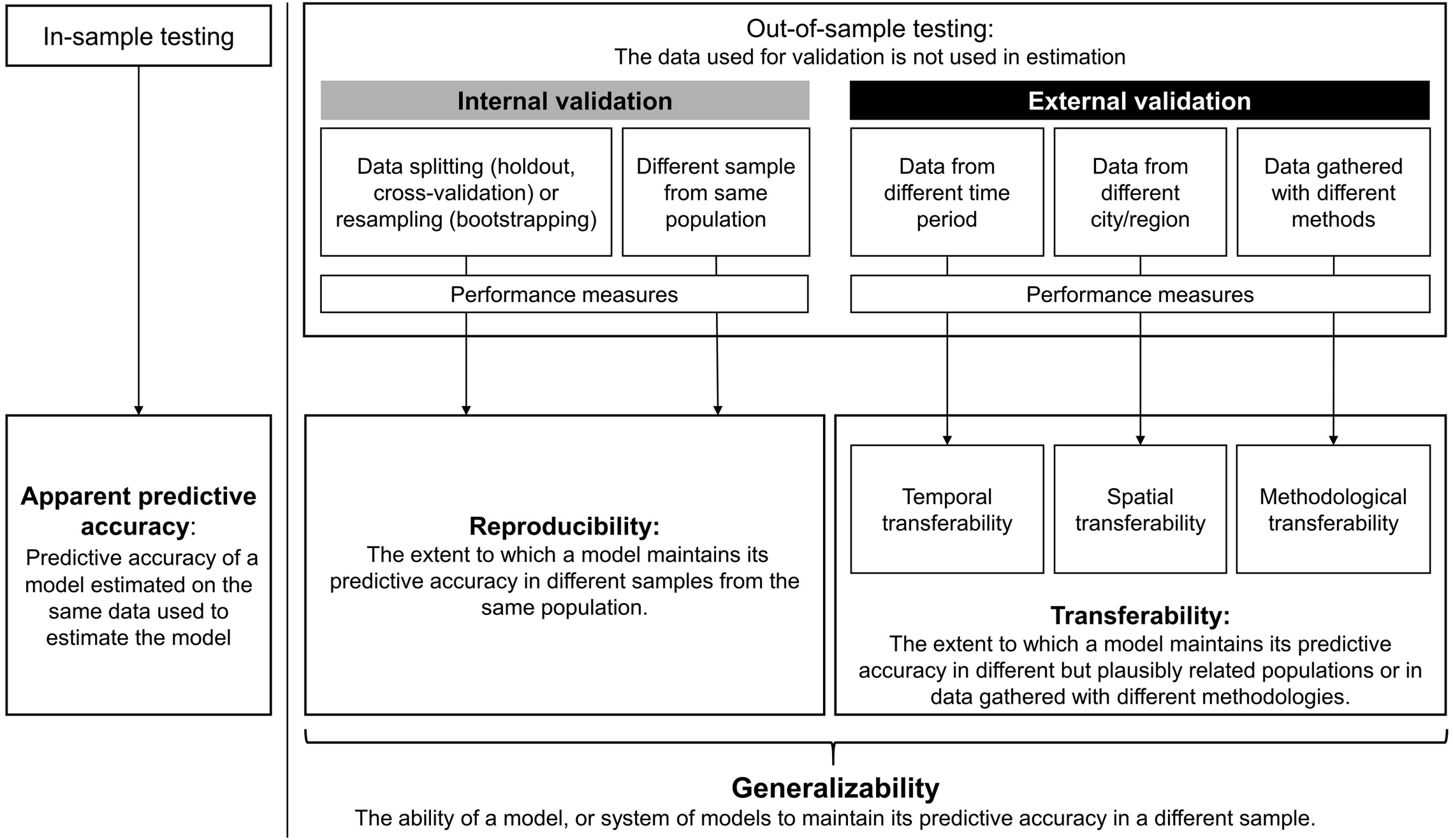
主要な用語定義

モデル検証（Model validation）：モデルに対する汎化性能の評価。

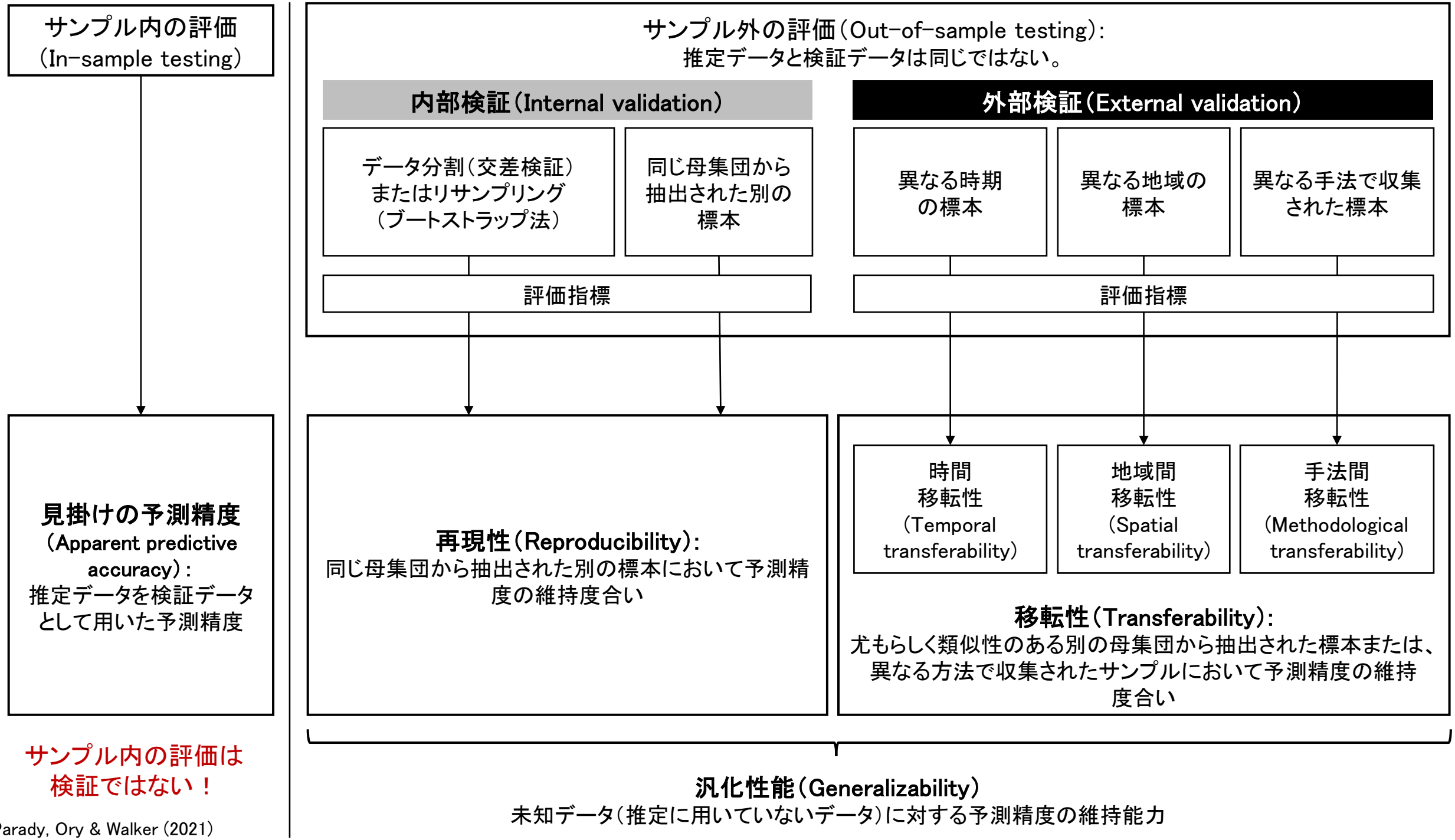
モデル検証の種類：

- **内部検証（Internal validation）：** 再現性の評価。
  - データ分割（Data splitting）
  - 同じ母集団から抽出された別の標本
- **外部検証（External validation）：**移転性の評価。
  - 時間移転性（Temporal transferability）
  - 地域間移転性（Spatial transferability）
  - 手法間移転性（Methodological transferability）

Parady, Ory & Walker (2021)



Parady, Ory & Walker (2021)



Parady, Ory & Walker (2021)

A brief introduction to internal validation (data splitting methods)

内部検証方法

データ分割法

- **ホールドアウト検証（Holdout validation）**：データを推定データと検証データに無作為に分割する。

2項アウトカムの場合、$\mathrm{Q}[y_n, \hat{y}_\mathrm{n}]$は個体$n$に対する予測の正確さの指標とする

$$\mathrm{Q}[y_n, \hat{y}_\mathrm{n}] = \begin{cases} 0 \; if \; y_n = \hat{y}_\mathrm{n} \\ 1 \; if \; y_n \neq \hat{y}_\mathrm{n} \end{cases}$$

ここで、y_nは個体nに対する観測されたアウトカムであり、$\hat{y}_\mathrm{n}$は個体nに対する予測されたアウトカムである。

ホールドアウトの推定量は次式のように定義する

$$HOV = \frac{1}{N_v} \sum_{n_v=1}^{N_v} Q[y_{n_v}, \hat{y}^e_{n_v}]$$

ここで、
y_{n_v}は検証データの個体nに対する観測されたアウトカムである。
$\hat{y}^e_{n_v}$は検証データの個体nに対する予測されたアウトカムである（推定データを用いたモデルによって予測されたもの）。

A brief introduction to internal validation (data splitting methods)

内部検証方法

データ分割法

- **交差検証（Cross-validation）:** データ分割をB回に行って、推定データと検証データB組を作成する。

  交差検証の推定量は次式のように定義する:

$$CV = \frac{1}{\mathrm{B}} \sum_{b} HOV_b$$

ここで、Bは推定と検証データ組の数である。

Parady, Ory & Walker (2021)

A brief introduction to internal validation (data splitting methods)

内部検証方法

交差検証：よく使われる交差検証法

$$CV = \frac{1}{\mathrm{B}} \sum_{b} HOV_b$$

推定データセットのサイスNeあらゆる可能なセットを用いる場合、exhaustive splittingと呼ばれる。　それ以外の場合は、partial splittingと呼ばれる（Arlot and Celisse, 2009）。

Partial splitting methods（より低い計算コスト）

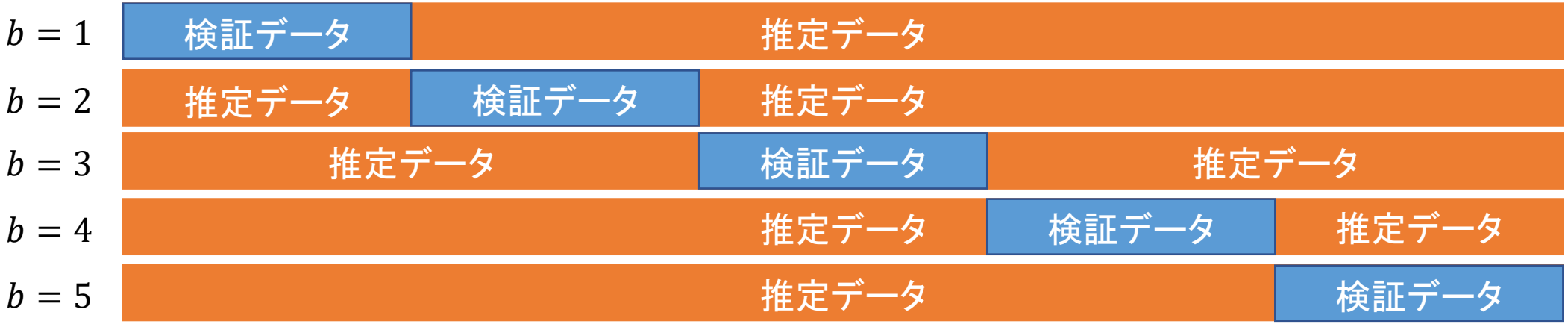
- **k−分割交差検証(K−fold cross−validation):** データを漏れなく、ダブりなく且つ、ほぼ同じサイスであるBサブセットに分割して、逐次的に違うサブセットを検証データとする。

- **繰り返しホールドアウト検証（Repeated learning−testing）:** ホールドアウトをB回に行う。

A brief introduction to internal validation (data splitting methods)

内部検証方法

交差検証：よく使われる交差検証法



5−分割交差検証のイメージ

繰り返しホールドアウト検証のイメージ

A brief introduction to internal validation (data splitting methods)

評価指標（Performance measures）

観測と予測マーケットシェアの比較

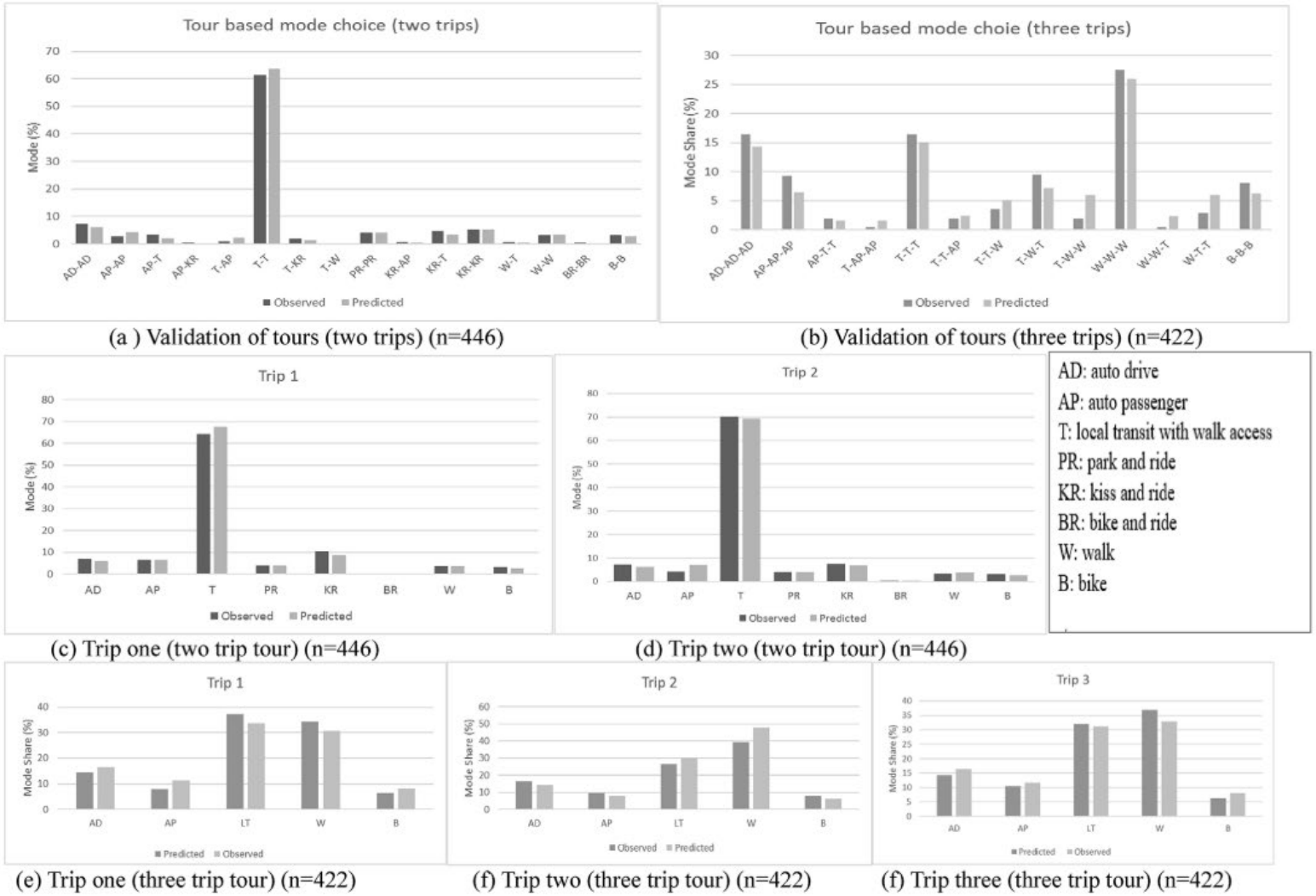


Fig. 3. Validation results of trips and tours.

Hasnine and Habib (2018)

A brief introduction to internal validation (data splitting methods)

評価指標（Performance measures）

的中率(Percentage of correct predictions): 選択確率の最も高い選択肢を予測選択とする。

ただし、

モデルAによる個人nに対する選択確率:

- **選択肢A:0.34＊**
- 選択肢B:0.33
- 選択肢C:0.33

モデルBによる個人nに対する選択確率:

- **選択肢A:0.50＊**
- 選択肢B:0.30
- 選択肢C:0.20

モデルCによる個人nに対する選択確率:

- **選択肢A:0.90＊**
- 選択肢B:0.05
- 選択肢C:0.05

＊実際に選ばれた選択肢

的中率は、モデル間の確率推定値の差を考慮しないため、これらのモデルは同等に評価される。

予測の明確さを考慮する指標が必要である。

A brief introduction to internal validation (data splitting methods)

評価指標（Performance measures）

予測の明確さ（Clearness of prediction）:

明確に正しい選択の割合（Percentage of clearly right choices）: 実際に選ばれた選択肢の選択確率が閾値 t を超えた割合である。

$$\%\mathrm{CR} = \frac{100}{N_v}\sum_{n_v=1}^{N_v} CR_{n_v} \qquad \text{ただし、} \qquad CR_{n_v} = \begin{cases} 1 \; if \; \hat{P}\left(y_{n_v}^{e}\right) > t \\ 0 \qquad otherwise \end{cases}$$

明確に違う選択の割合（Percentage of clearly wrong choices）:選ばれなかった選択肢の選択確率が閾値 t を超えた割合である。

$$\%\mathrm{CW} = \frac{100}{N_v}\sum_{n_v=1}^{N_v} \mathrm{CW}_{n_v} \qquad \text{ただし、} \qquad CW_{n_v} = \begin{cases} 1 \; if \; \hat{P}\left(!\, y_{n_v}^{e}\right) > t \\ 0 \qquad otherwise \end{cases}$$

ここで、$\hat{P}\left(!\, y_{n_v}^{e}\right)$ は実際に選ばれた選択肢以外の選択肢の選択確率である。

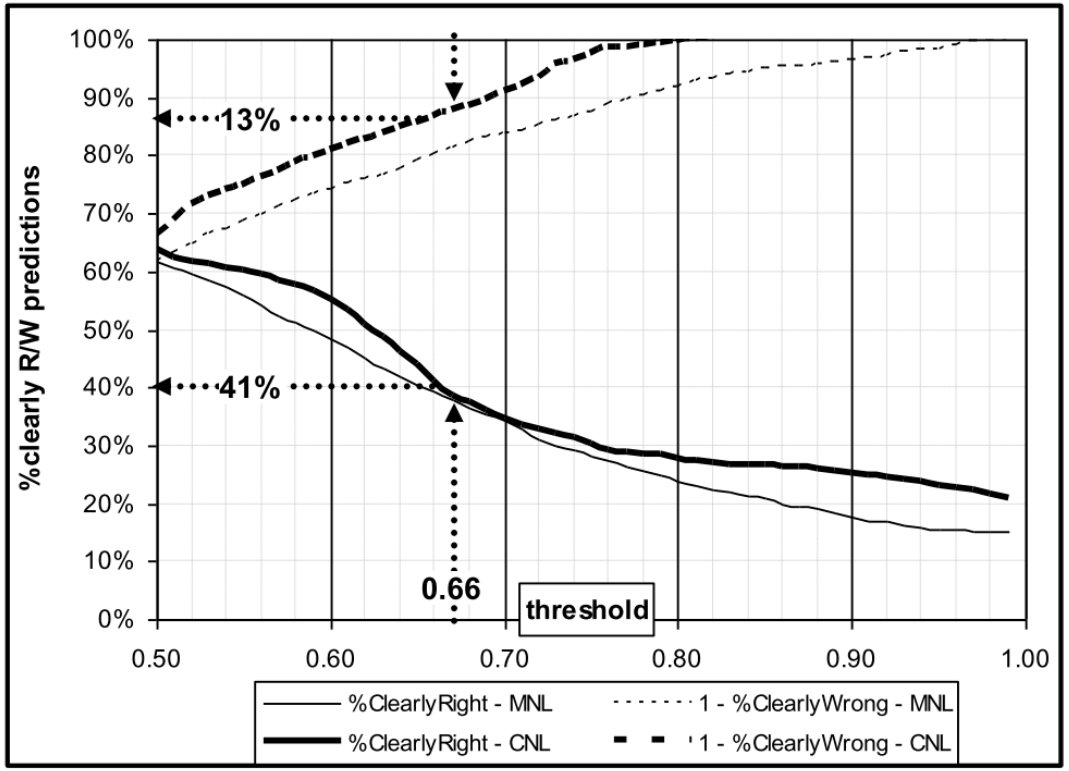
de Luca and Cantarella (2009)

A brief introduction to internal validation (data splitting methods)

評価指標（Performance measures）

予測の明確さ（Clearness of prediction）：閾値tをどう決めるか

- 有意義な評価指標になるため、tは c^{-1}より十分大きいでなければならない（ cは選択肢集合のサイズである）。
- 既往研究で使われた値：
  - 2択モデル：$t = 0.9$（de Luca and Di Pace, 2015）
  - 3択モデル：$t = 0.5$（Glerum, Atasoy and Bierlaire , 2014）



de Luca and Di Pace (2015)

See appendix for a list of commonly used indicators

Validation and reporting practices in the transportation academic literature

交通分野におけるモデル検証の報告状況

226 偏をレビューした

92% の論文が適合度指標を報告した

64.6% の論文が政策に関連する推論を報告した

限界効果、弾力性、オッズ比、時間価値推定値、シナリオシミュレーション等。

18.1% の論文がモデル検証を報告した。

文献において利用された内部検証方法

Internal validation methods reported in the literature by frequency.

Method	Abbvr.	Frequency	Percentage
Holdout validation	HOV	18	56.3%
Repeated learning-testing	RLT	8	25.0%
Validation against an independent sample	IS	4	12.5%
Repeated K-fold cross-validation	R-K-CV	1	3.1%
Other sample splitting methods	SS-O	1	3.1%

Towards better validation practices in the field

現状を改善するため

- **モデル検証を義務化する**
  - 査読付学術雑誌において義務化する。
  - 機械学習研究において、交差検証は必須である。
- **ベンチマークデータを共有する**
  - 交通分野において、ベンチマークデータとコードを共有する文化がない（最近少し変わっているが）。
- **検証研究を促進する**
  - 新規性を過剰に求められる。
  - 既存のモデルや理論を検証する投稿を促進すべきである。
- **結果報告ガイドラインを作成する**

ちょっとまって。。。

“I’m not validating my model because I’m not trying to build a predictive framework. I’m trying to learn about travel behavior”

「予測モデルを構築するつもりではなく、交通行動について学びたいだけだ」

従来分析に近ければ近いほど（例：手段選択、目的地選択、経路選択等）、検証を行う責任がある。

ちょっとまって。。。

“Should every study using a discrete choice model be conducting validation?”

「離散選択モデルを用いる研究すべてが検証を行うべきか」

横断的な観察研究への依存と、アカデミアにおいてフィードバックのなさによる、少なくとも、政策示唆することであれば検証を行うべきである。

発表される研究成果を疑う理由は少なくはない（むしろ、疑うべきである）。最も当たり前のは過剰適合(overfitting)の問題である。

ちょっとまって。。。

“Is what we learn about travel behavior from coefficient estimation less valuable if not conducted?”

「検証を行わないと、係数推定から学ぶことの価値が下がるか」

推定係数は政策に関する推論に役立つが、係数そのものからモデルの予測精度についてなにも分からない。政策に関する推論と共にその推論の汎化性能を報告すべきである。

Finally

検証を行うだけで、再現性の危機が解決しないが、正しい方向への一歩である。

モデル検証は因果関係問題の解決法でもない。ただし、横断的な観察研究への依存に対して、研究成果の汎化性能を改善するために不可欠である。

References:

1. Ben-Akiva, M. E., Lerman, S. R. (1985). Discrete choice analysis: theory and application to travel demand. MIT press.

2. M. Baker, D. Penny (2016) Is there a reproducibility crisis? Nature, 533 (7604) pp. 452-454

3. de Luca, S. De and Cantarella, G. E. (2009) 'Validation and comparison of choice models', in Saleh, W. and Sammer, G. (eds) Travel Demand Management and Road User Pricing: Success, Failure and Feasibility. Ashgate publications, pp. 37–58. doi: 10.1017/cbo9780511619960.008.

4. de Luca, S., and R. Di Pace (2015). Modelling Users' Behaviour in Inter-Urban Carsharing Program: A Stated Preference Approach. Transportation Research Part A: Policy and Practice, Vol. 71, pp. 59–76. https://doi.org/10.1016/j.tra.2014.11.001.

5. Glerum, A., Atasoy, B. and Bierlaire, M. (2014) 'Using semi-open questions to integrate perceptions in choice models', Journal of Choice Modelling. Elsevier, 10(1), pp. 11–33. doi: 10.1016/j.jocm.2013.12.001.

6. Hasnine, M. S. and Habib, K. N. (2018) 'What about the dynamics in daily travel mode choices? A dynamic discrete choice approach for tour-based mode choice modelling', Transport Policy. Elsevier Ltd, 71(August), pp. 70–80. doi: 10.1016/j.tranpol.2018.07.011.

7. Hensher, D. A., Rose, J. M., & Greene, W. H. (2015). Applied choice analysis: a primer. Cambridge University Press. 2nd Edition.

8. Kamargianni, M., M. Ben-Akiva, and A. Polydoropoulou (2014) Incorporating Social Interaction into Hybrid Choice Models. Transportation, Vol. 41, No. 6, pp. 1263–1285.

9. Qin, H., J. Gao, H. Guan, and H. Chi. (2017) Estimating Heterogeneity of Car Travelers on Mode Shifting Behavior Based on Discrete Choice Models. Transportation Planning and Technology, Vol. 40, No. 8, pp. 914–927. https://doi.org/10.1080/03081060.2017.1355886.

10. Khan, M., K. M. Kockelman, and X. Xiong. (2014) Models for Anticipating Non-Motorized Travel Choices, and the Role of the Built Environment. Transport Policy, Vol. 35, pp. 117–126. https://doi.org/10.1016/j.tranpol.2014.05.008.

11. McCloskey, D. N., and S. T. Ziliak.(1996) The Standard Error of Regressions. Journal of Economic Literature, Vol. 34, No. 1, pp. 97–114.

12. Parady, G., Ory, D., Walker, J. (2021) The overreliance on statistical goodness of fit and under-reliance on validation in discrete choice models: A review of validation practices in the transportation academic literature . Journal of Choice Modelling 38, 100257 (Open Access)

13. Parady, G., Axhausen K.W.: Size Matters (2023) The Use and Misuse of Statistical Significance in Discrete Choice Models in the Transportation Academic Literature. Transportation (accepted)

Appendix: Definition of model validation performance measures reported in the literature

Parady, Ory & Walker (2021)

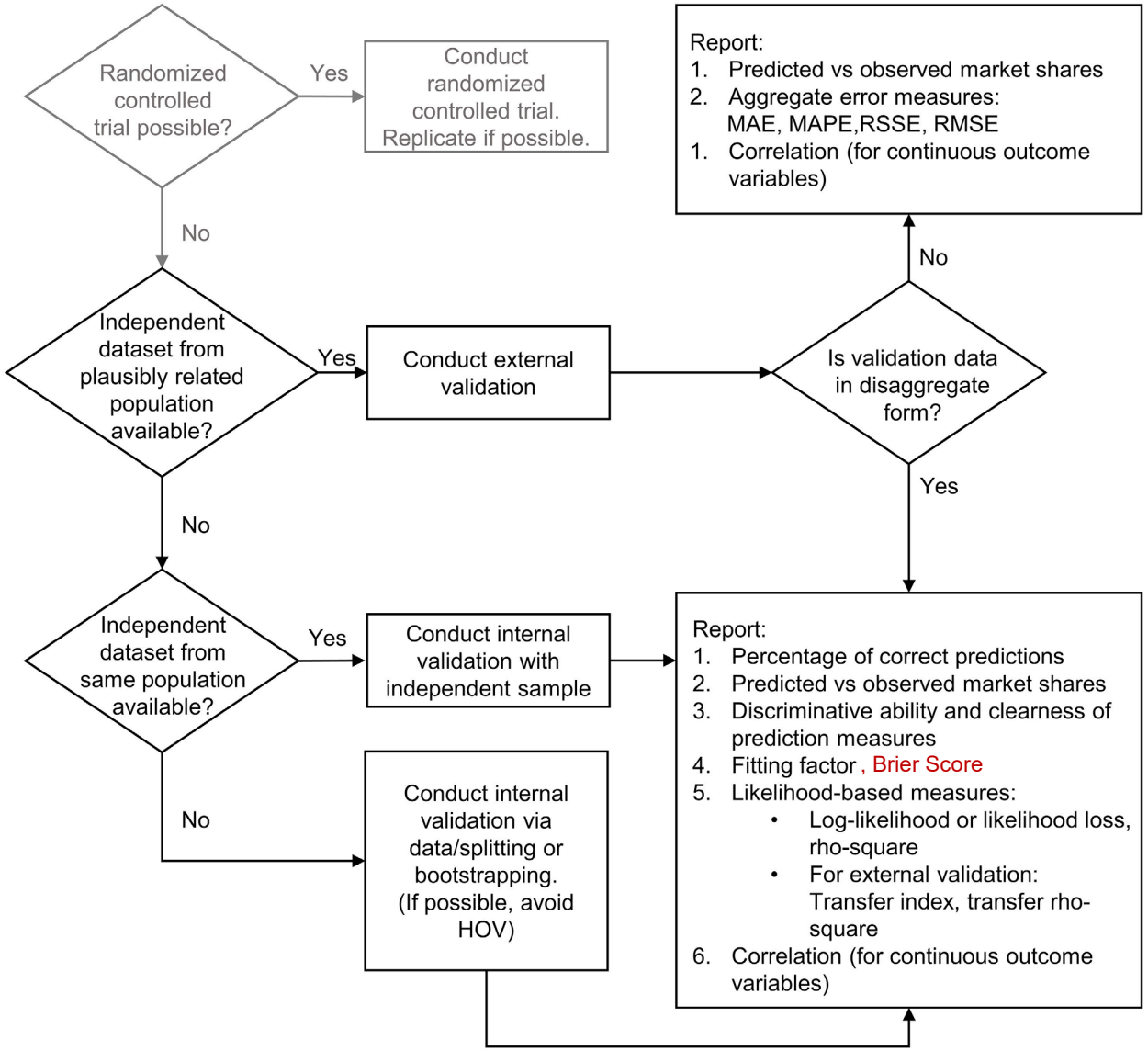
Index		Type	Formula	Notes
Mean absolute percentage error 平均絶対誤差率	MAPE	*Absolute*	$\frac{100}{M}\sum_{m=1}^{M}\left\|\frac{\hat{s}^{e}{}_{v,m}-s_{v,m}}{s_{v,m}}\right\|$	M is the number of alternatives in the choice set. $s_{v,m}$ is an aggregate outcome measure in sample *v*, such as the market share of alternative *m* (i.e. modal market share), choice frequency, etc. $\hat{s}^{e}{}_{v,m}$ is an aggregate outcome measure in sample *v*, such as the market share of alternative *m*, predicted from model estimated on sample *e*. $\hat{P}(y^{e}_{n_v,m})$ is the predicted probability that individual *n* chooses alternative *m,* predicted from model estimated on sample *e*. y_{nm} is the actual outcome variable valued 0 or 1.
Root sum of square error 二乗平方根誤差和	RSSE	Relative	$\sqrt{\sum_{m=1}^{M}\left(\hat{s}^{e}{}_{v,m}-s_{v,m}\right)^{2}}$	
Mean absolute error 平均絶対誤差	MAE	Aggregate: Relative Disaggregate: Absolute	$\frac{1}{M}\sum_{m=1}^{M}\left\|\hat{s}^{e}{}_{v,m}-s_{v,m}\right\|$	
Mean squared error 平均二乗誤差	MSE	Aggregate: Relative Disaggregate: Absolute	$\frac{1}{M}\sum_{m=1}^{M}\left(\hat{s}^{e}{}_{v,m}-s_{v,m}\right)^{2}$	
Root mean square error 二乗平均平方根誤差	RMSE	Aggregate: Relative Disaggregate: Absolute	$\sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{s}^{e}{}_{v,m}-s_{mv,}\right)^{2}}$	
Brier Score ブライアスコア	BS	Absolute	$\frac{1}{N_v}\sum_{n_v=1}^{N_v}\sum_{m=1}^{M}\left(\hat{P}(y^{e}_{n_v,m})-y_{n_v,m}\right)^{2}$	

Appendix: Definition of model validation performance measures reported in the literature

Parady, Ory & Walker (2021)

Index	Type		Formula	Notes
Log-likelihood 対数尤度	LL	Relative	$LL_v(\hat{\boldsymbol{\beta}}^e)$	$LL_{v,r}(\hat{\boldsymbol{\beta}}^e)$ is log-likelihood of the model estimated on data *e* applied to the validation data v_r. $N_{v,r}$ is the size of the validation (holdout) sample r, and R is number of validation samples generated. $LL_v(\mathbf{0})$ is log-likelihood of the model when all parameters are zero for data *v*. $LL_v(\hat{\boldsymbol{\beta}}^v)$ is the likelihood of the model estimated on the validation data *v*. $LL_v(\boldsymbol{MS}^v)$ is a base model estimated on validation data *v* (i.e. market share model.) ρ^2_{local} is the local rho-square of the model. f_m is the observed choice frequency of alternative m in sample *v*, and $E(f^e_{v,m})$ is the expected choice frequency predicted from model estimated on sample *e*.
Log-likelihood loss 対数尤度損失	LLL	Absolute	$\frac{1}{\mathrm{R}}\sum_r -\frac{1}{N_{v,r}}\sum_{n_{v,r}} LL_{v,r}(\hat{\boldsymbol{\beta}}^e)$ $\forall\ 1 \le r\ \le R$	
Rho-square σ^2	RHOSQ	Absolute	$\rho^2 = 1 - \frac{LL_v(\hat{\boldsymbol{\beta}}^e)}{LL_v(\mathbf{0})}$	
Transfer rho-square 移転 σ^2	T-RHOSQ	Relative	$\rho^2_{transfer} = 1 - \frac{LL_v(\hat{\boldsymbol{\beta}}^e)}{LL_v(\boldsymbol{MS}^v)}$	
Transfer index 移転指標	TI	Pass/Fail	$\frac{LL_v(\hat{\boldsymbol{\beta}}^e) - LL_v(\boldsymbol{MS}^v)}{LL_v(\hat{\boldsymbol{\beta}}^v) - LL_v(\boldsymbol{MS}^v)}$	
Transferability test statistic 移転性検定統計量	TTS	Relative	$-2\left(LL_v(\hat{\boldsymbol{\beta}}^v) - LL_v(\hat{\boldsymbol{\beta}}^e)\right)$	
χ^2 **test**	CHISQ	Pass/Fail	$\sum_{m=1}^{M} \frac{\left(f_m - E(f^e_{v,m})\right)^2}{E(f^e_{v,m})}$	

Appendix: Validation and reporting practices in the transportation academic literature



Heuristic to select validation method given available resources and recommended performance measures to report

Appendix: Validation and reporting practices in the transportation academic literature

Table 4

Predictive accuracy performance measures reported in the literature by frequency.

Performance measure	Abbrv.	Frequency	Percentage
Log-likelihood/log-likelihood loss	LL/LLL	19	46.3%
Percentage of correct predictions or First Preference Recovery	FPR	10	24.4%
Predicted vs observed market outcomes	PVO	10	24.4%
Mean absolute error	MAE	6	14.6%
Root mean square error	RMSE	4	9.8%
Error/Percentage error/Absolute percentage error	E/PE/APE	3	7.3%
Rho-Square	RHOSQ	3	7.3%
Transfer index	TI	2	4.9%
% clearly right (t)	%CR	1	2.4%
Brier Score	BS	1	2.4%
Chi-square	CHISQ	1	2.4%
Concordance index	C	1	2.4%
Correlation	CORR	1	2.4%
Fitting factor	FF	1	2.4%
Mean absolute percentage error	MAPE	1	2.4%
Sum of square error	SSE	1	2.4%
Transferability test statistic	TTS	1	2.4%
All other measures specified in Table 1	–	0	0%
Other measures not specified in Table 1	–	3	7.3%

Very similar measures are reported jointly.