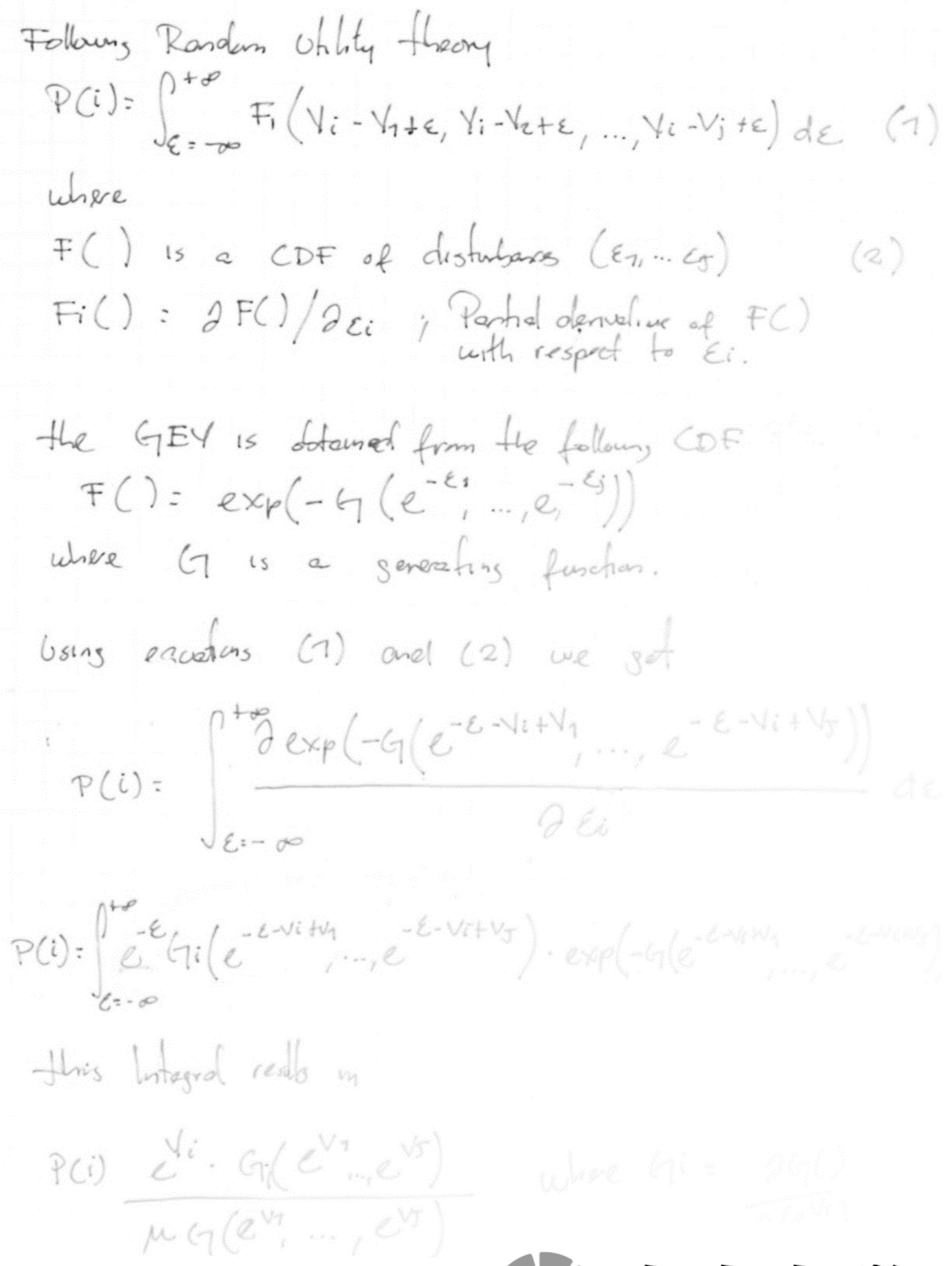
The 19th Behavior Modeling Summer School

Sep. 12 – 15 , 2020 @ The University of Tokyo



Basic inference and validation in discrete choice modeling

Giancarlos Troncoso Parady – The University of Tokyo

Basic inference discrete choice modeling

Following Random utility theory

$$P(i) = \int_{\varepsilon=-\infty}^{+\infty} F_i(V_i - V_1 + \varepsilon, V_i - V_2 + \varepsilon, \ldots, V_i - V_j + \varepsilon)\, d\varepsilon \quad (1)$$

where

$F(\,)$ is a CDF of disturbances $(\varepsilon_1, \ldots, \varepsilon_J)$ (2)

$F_i(\,) : \partial F(\,) / \partial \varepsilon_i$; Partial derivative of $F(\,)$ with respect to ε_i.

The GEV is obtained from the following CDF

$$F(\,) = \exp(-G(e^{-\varepsilon_1}, \ldots, e^{-\varepsilon_J}))$$

where G is a generating function.

Using equations (1) and (2) we get

$$P(i) = \int_{\varepsilon=-\infty}^{+\infty} \frac{\partial \exp(-G(e^{-\varepsilon - V_i + V_1}, \ldots, e^{-\varepsilon - V_i + V_J}))}{\partial \varepsilon_i} d\varepsilon$$

$$P(i) = \int_{\varepsilon=-\infty}^{+\infty} e^{-\varepsilon} G_i(e^{-\varepsilon - V_i + V_1}, \ldots, e^{-\varepsilon - V_i + V_J}) \cdot \exp(-G(e^{-\varepsilon - V_i + V_1}, \ldots, e^{-\varepsilon - V_i + V_J}))$$

This integral results in

$$P(i) = \frac{e^{V_i} \cdot G_i(e^{V_1}, \ldots, e^{V_J})}{\mu G(e^{V_1}, \ldots, e^{V_J})} \quad \text{where } G_i = \partial G(\,) / \ldots$$

東京大学 THE UNIVERSITY OF TOKYO

Why is inference important ?

Variable name	Coefficient	S.E.	t statistic
Auto constant	1.45	0.393	3.70
In-vehicle time (min)	-0.0089	0.0063	-1.42
Out-of-vehicle time (min)	-0.0308	0.0106	-2.90
Auto out-of-pocket cost (c)	-0.0115	0.0026	-4.39
Transit fare	-0.0070	0.0038	-1.87
Auto ownership (specific to auto mode)	-0.770	0.213	3.16
Downtown workplace (specific to auto mode)	-0.561	0.306	-1.84
Number of observations	1476		
Number of cases	1476		
LL(0)	-1023		
LL(β)	-347.4		
-2[LL(0)-LL(β)]	1371		
ρ^2	0.660		
$\bar{\rho}^2$	0.654		

Table adapted from Ben-Akiva and Lerman (1985)

← Magnitudes are not directly interpretable. We can only interpret the effect direction, or use them to calculate utilities, and choice probabilities

To make some sense of these parameters we must calculate elasticities or marginal effects

Basic Inference in discrete choice models

MNL: Logit Elasticities (Point elasticities)

- **Direct elasticity:** measures the **percentage change in the probability** of choosing a particular alternative in the choice set with respect to a given **percentage change** in an attribute of that same alternative.

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

- **Cross-elasticity:** measures the **percentage change in the probability** of choosing a particular alternative in the choice set with respect to a given **percentage change** in a competing alternative.

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

← Because of IIA, cross-elasticities are uniform across all alternatives

Hensher, David A., John M. Rose, and William H. Greene (2015)

Basic Inference in discrete choice models

MNL: Logit Elasticities (Point elasticities)

- The elasticities shown before are **individual elasticities (Disaggregate)**
- To calculate sample (aggregate) elasticities we use the **probability weighted sample enumeration** method:

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

Sample direct elasticity

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

Sample cross-elasticity

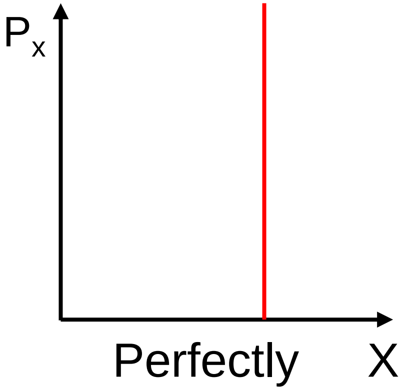
Where $\overline{P(i)}$ is the aggregate choice probability of alternative I, and $\hat{P}_{in}(i)$ is an estimated choice probability

- Uniform cross-elasticities do not necessarily hold at the aggregate level
- Also note that elasticities for dummy variables are **meaningless!**

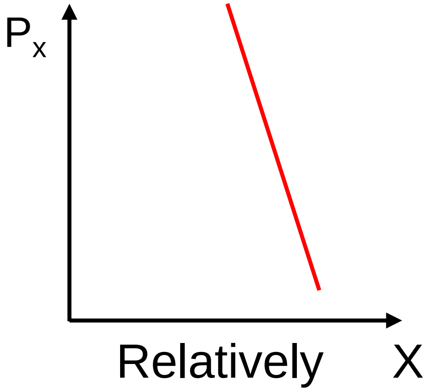
Hensher, Rose, and Greene (2015)

Basic Inference in discrete choice models

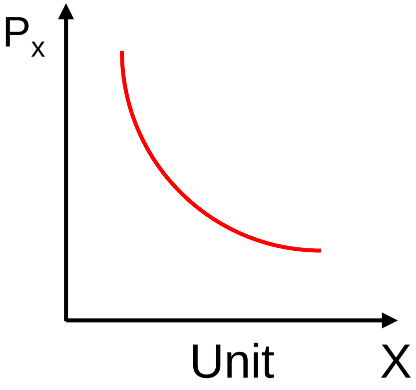
Relation between elasticity of demand, change in price and revenue



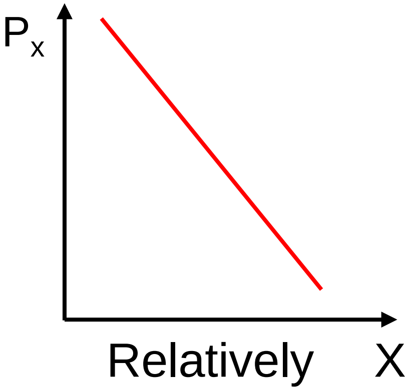
Perfectly inelastic



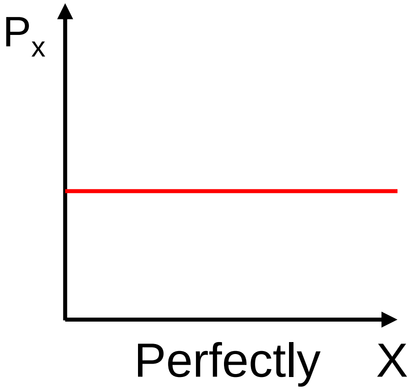
Relatively inelastic



Unit elastic



Relatively elastic



Perfectly elastic

Direct elasticity:

1% increase in X results in a 0% decrease in P(i)

1% increase in X results in a less than 1% decrease in P(i)

1% increase in X results in a 1% decrease in P(i)

1% increase in X results in a more than 1% decrease in P(i)

1% increase in X results in a ∞ decrease in P(i)

Cross elasticity:

1% increase in X results in a 0% increase in P(j)

1% increase in X results in a less than 1% increase in P(j)

1% increase in X results in no percent change in P(j)

1% increase in X results in a more than 1% increase in P(j)

1% increase in X results in a ∞ increase in P(j)

Hensher, Rose, and Greene (2015)

Basic Inference in discrete choice models

MNL: Marginal Effects

- **Direct marginal effects:** measures the **change in the probability** (absolute change) of choosing a particular alternative in the choice set with respect to a **unit change** in an attribute of that same alternative.

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

- **Cross-marginal effects:** measures the **change in the probability** (absolute change) of choosing a particular alternative in the choice set with respect to a **unit change** in a competing alternative.

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

Hensher, Rose, and Greene (2015)

Basic Inference in discrete choice models

MNL: Marginal Effects

- We can also calculate sample (aggregate) marginal effects we using e the **probability weighted sample enumeration** method:

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

Sample direct marginal effect

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

Sample cross-marginal effect

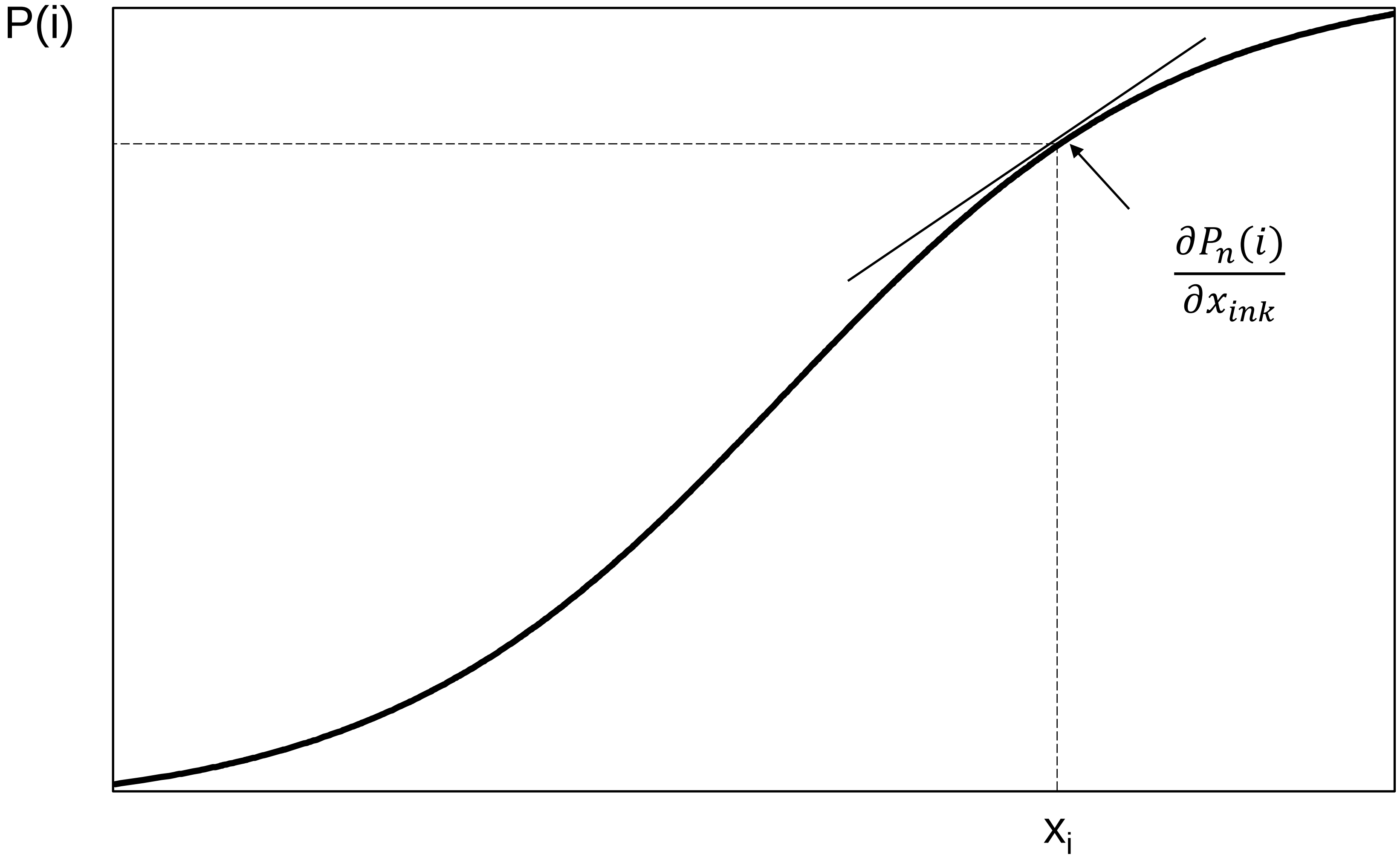
Where $\overline{P(i)}$ is the aggregate choice probability of alternative I, and $\hat{P}_{in}(i)$ is an estimated choice probability

- Marginal effects for dummy variables **do make sense** as we are talking about unit changes. But you still need to frame the problem correctly.

Hensher, Rose, and Greene (2015)

Basic Inference in discrete choice models

MNL: Marginal Effects



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

Hensher, David A., John M. Rose, and William H. Greene (2015)

Basic Inference in discrete choice models

Incremental Logit for prediction

- An alternative approach to using elasticities or marginal effects for prediction
- Prediction of changes in behavior based on existing choice probabilities

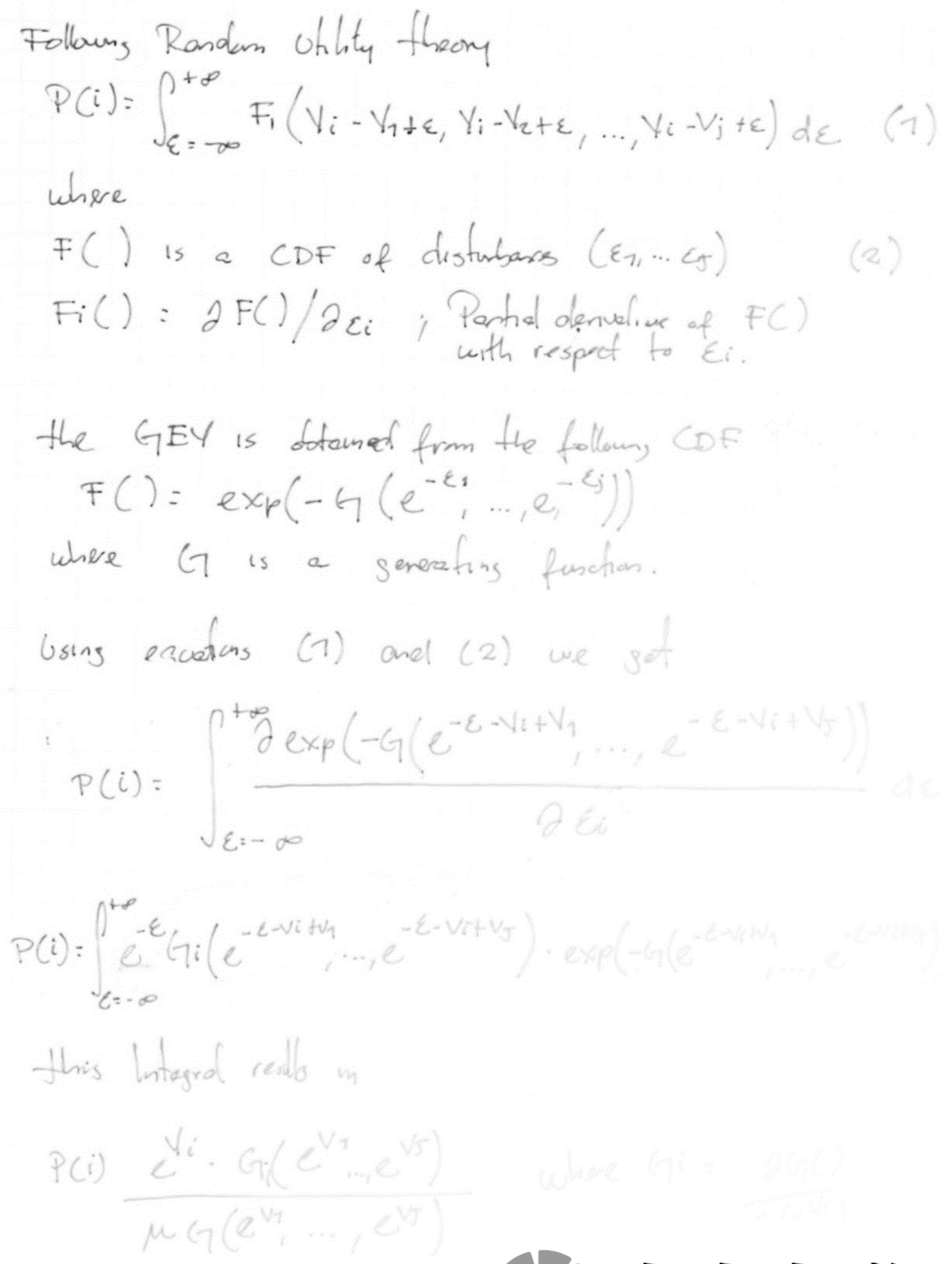
$$P'(i) = \frac{\exp(V_{in} + \Delta V_{in})}{\sum_{j \in C} \exp\left(V_{jn} + \Delta V_{jn}\right)}; \qquad where\ \Delta V_{in} = \sum_{k=1}^{K} \beta_k \Delta x_{ink}$$

Δx_{ink} *is a marginal change in the kth independent variable for alternative i and individual n*

- In fact, for linear-in-parameter models we need not calculate the utilities again

$$P'(i) = \frac{\exp(V_{in} + \Delta V_{in})}{\sum_{j \in C} \exp\left(V_{jn} + \Delta V_{jn}\right)} = \frac{\mathrm{P}(i)\exp(\Delta V_{in})}{\sum_{j \in C} P(j) \exp\left(\Delta V_{jn}\right)}$$

Ben-Akiva and Lerman (1985)



Validation practices in discrete choice modeling

東京大学
THE UNIVERSITY OF TOKYO

A credibility crisis in science and engineering?

Most published research findings are likely to be false due to factors such as lack of power of the study, small effect sizes, and great flexibility in research design, definitions, outcomes and methods.

(Ioannidis, 2005)

A credibility crisis in science and engineering?

What about the transportation field?

In practice:

Demand forecasting is the **"Achilles' heel"** of the transport planning model (Banister, 2002)

Demand overestimation: **30% Highway trips**

35% Transit trips (UK, 1962 – 1972)

→ **Forecasts have not become more accurate (between 1969-1998).** (Flyvbjerg, kamris Holm, & Buhl, 2005)

Unlike the natural sciences

- Dependence on cross-section observational studies.
- Classic scientific hypothesis testing is more difficult.
- Underscores the need for proper validation practices.

> *"There is little tradition of confronting and confirming predictions of cross-sectional models with outcomes in either back-casting or detailed before-and-after studies"*
>
> (Boyce & William, 2015)

→ **While in practice, a feedback loop exists between forecast outputs and implementation results in the form of measurable forecasting errors, in academia such feedback loop rarely exists.**

Term definitions and research scope

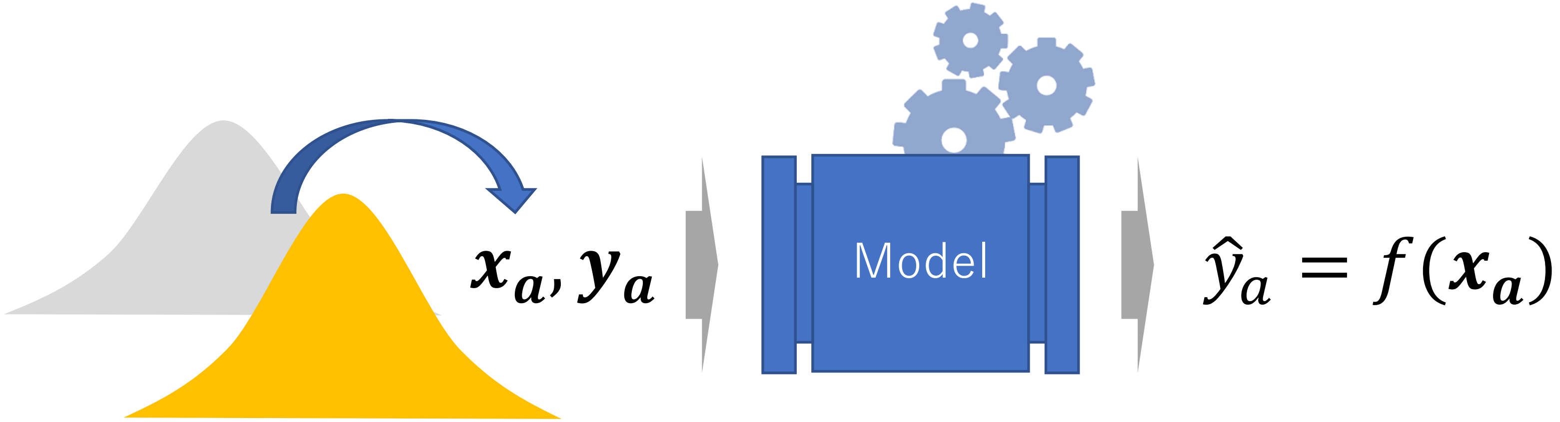
- ***Predictive accuracy*:** the degree to which predicted outcomes match observed outcomes.
    - ***Calibration*:*** the agreement between observed outcomes and predictions
    - ***discrimination:*** the ability of model predictions to discriminate between those with and those without an observed outcome
- ***Generalizability:*** the ability of a model or system of models to maintain its predictive accuracy in a different sample.
    - ***Reproducibility*:** the ability of a model or system of models to maintain its predictive accuracy in different samples from the same population.
    - ***Transportability (transferability)*** the ability of a model or system of models to maintain its predictive accuracy in samples from different but plausibly related populations or in samples collected with different methodologies.

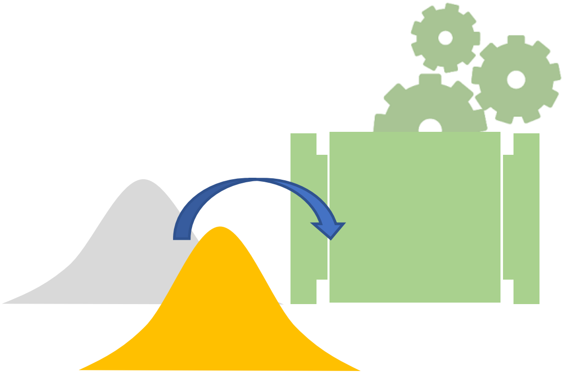
Term definitions and research scope

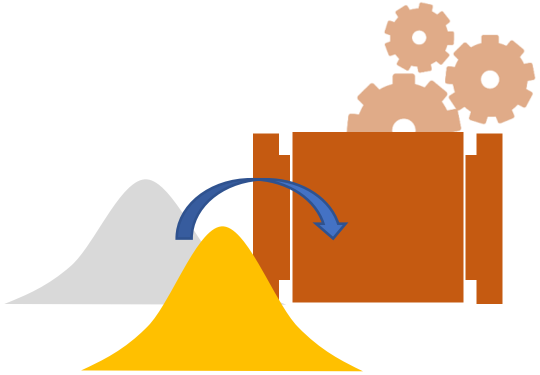
- ***Validation*:** the evaluation of the generalizability of a statistical model.
    - **Internal validation** evaluates **reproducibility**
    - **External validation** evaluates **transferability**

A general overview of model validation methods

Estimation

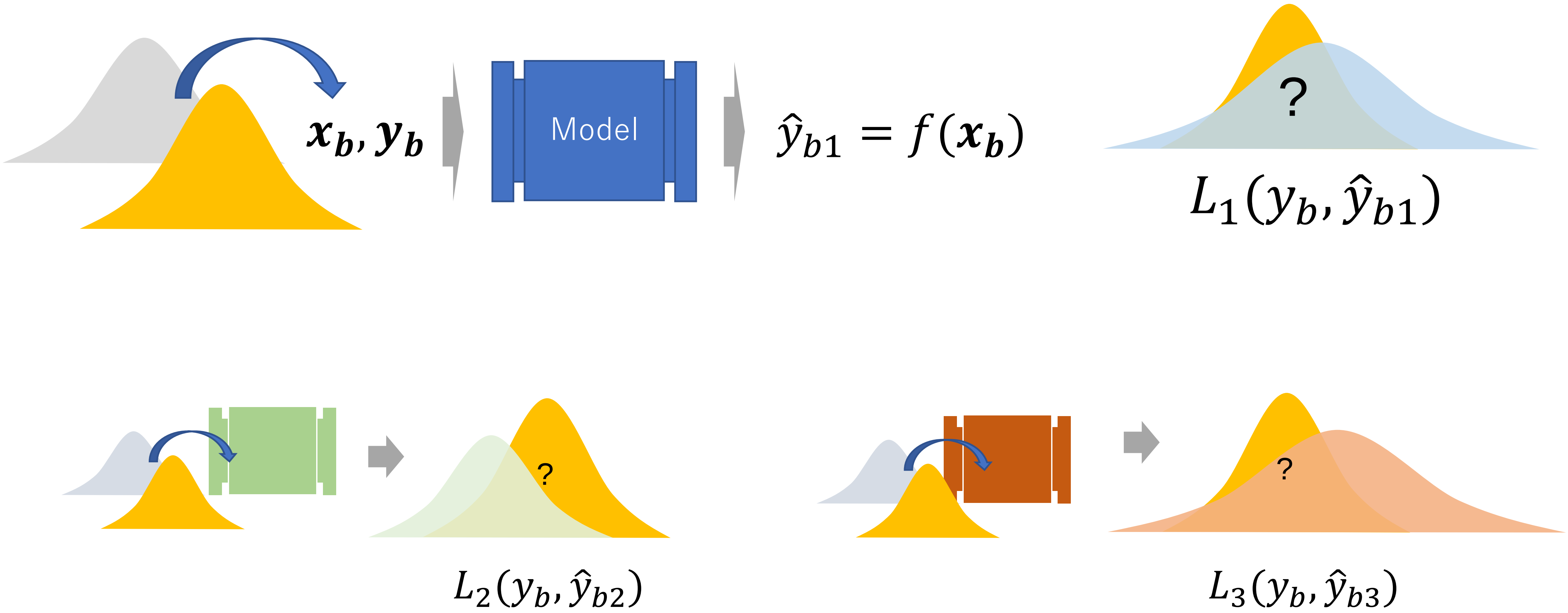


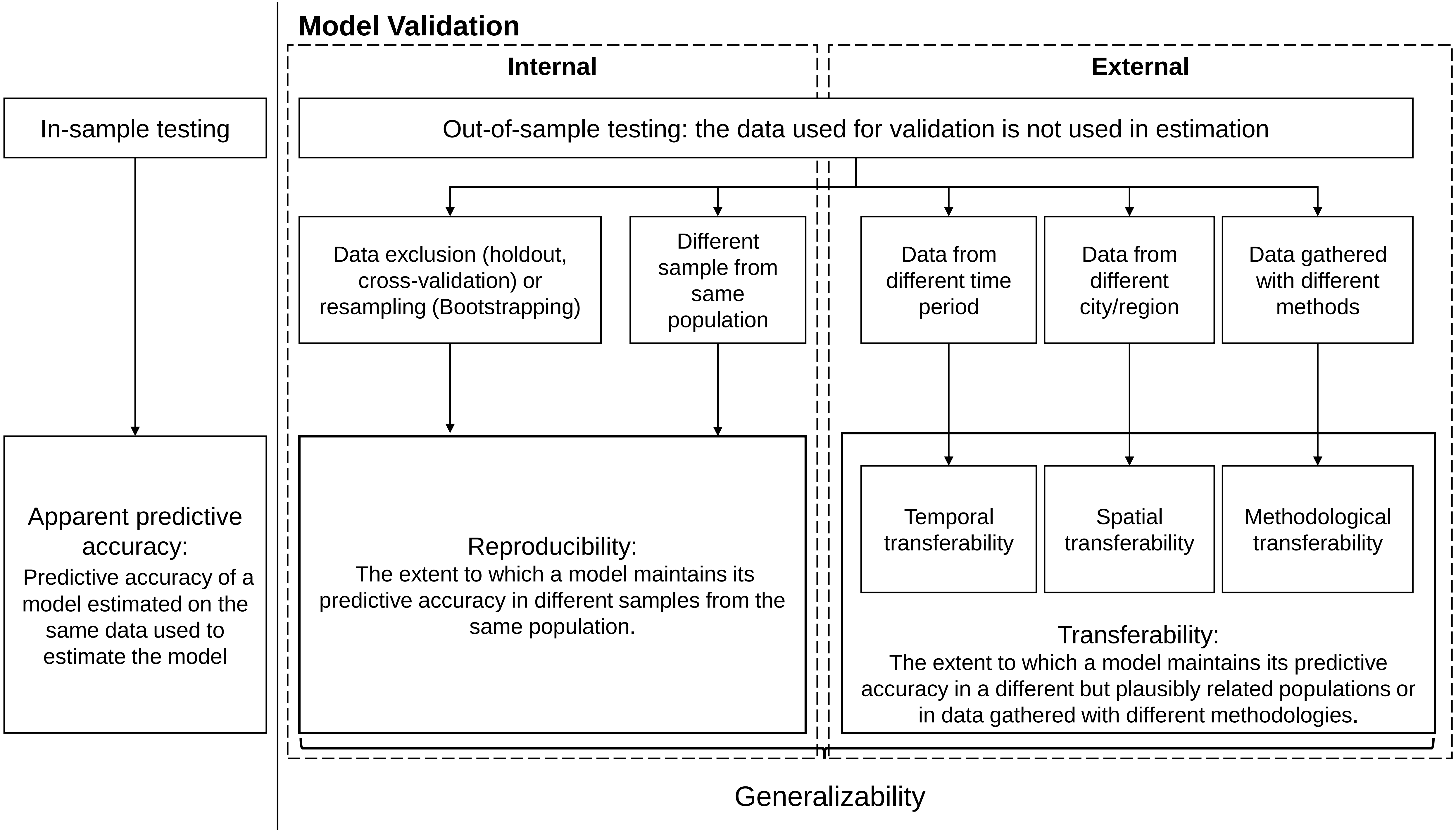




A general overview of model validation methods

Validation




Model Validation
Internal
External
In-sample testing
Out-of-sample testing: the data used for validation is not used in estimation
Data exclusion (holdout, cross-validation) or resampling (Bootstrapping)
Different sample from same population
Data from different time period
Data from different city/region
Data gathered with different methods
Apparent predictive accuracy:
Predictive accuracy of a model estimated on the same data used to estimate the model
Reproducibility:
The extent to which a model maintains its predictive accuracy in different samples from the same population.
Temporal transferability
Spatial transferability
Methodological transferability
Transferability:
The extent to which a model maintains its predictive accuracy in a different but plausibly related populations or in data gathered with different methodologies.
Generalizability

A general overview of model validation methods

(Some) Performance measures

Direct prediction accuracy measures

- Directly interpretable
- Allows for comparison across models (to some extent)

Relative prediction accuracy measures

- Scores not directly interpretable
- Only meaningful in relative terms
- Useful for model selection

The best model among a set of models can still be a very bad model

Measure	Abbrv.
Predicted vs observed market outcomes	PVO
Percentage of correct predictions	FPR
% clearly right (t)	%CR
% clearly wrong (t)	%CW
% unclear (t)	%U
Fitting factor	FF
Correlation	Corr
ρ^2	RHOSQ
Brier Score	BS
Absolute percentage error	APE
Sum of square error	SSE
Root sum of square error	RSSE
Mean absolute error	MAE
Mean absolute percentage error	MAPE
Mean squared error	MSE
Root mean square error	RMSE
χ^2 test*	CHISQ
Log-likelihood	LL
Mean log-likelihood loss	MLLL
likelihood ratio test (LR), AIC	f(LL)

See Parady, Ory and Walker (2019) for specific details

Validation and reporting practices in the transportation literature

Using the Web of Science Core Collection maintained by Clarivate Analytics we reviewed validation and reporting practices in the transportation literature from the last 5 years (2014 to 2018). Articles were selected based on the following criteria:

- Peer-reviewed journal articles published between **2014 and 2018**
- Analysis uses **discrete choice models**
- Target choice dimensions are:

  Destination choice, Model choice and Route choice
- Web of Science Database fields are:

  Transportation; transportation science and technology; economics; civil engineering
- Research scope is limited to **land transport and daily travel behavior** (tourism, evacuation behavior, etc. were excluded)
- Articles use **empirical data** (Studies using numerical simulations only were excluded)
- Methodological papers only included if the use empirical data
- SP only papers are excluded

Validation and reporting practices in the transportation academic literature

228 articles reviewed

92% reported a goodness of fit statistics

64% reported a policy-related inference

Marginal effects, elasticities, odds ratios, value of time estimates, marginal rates of substitution, and policy scenario simulations

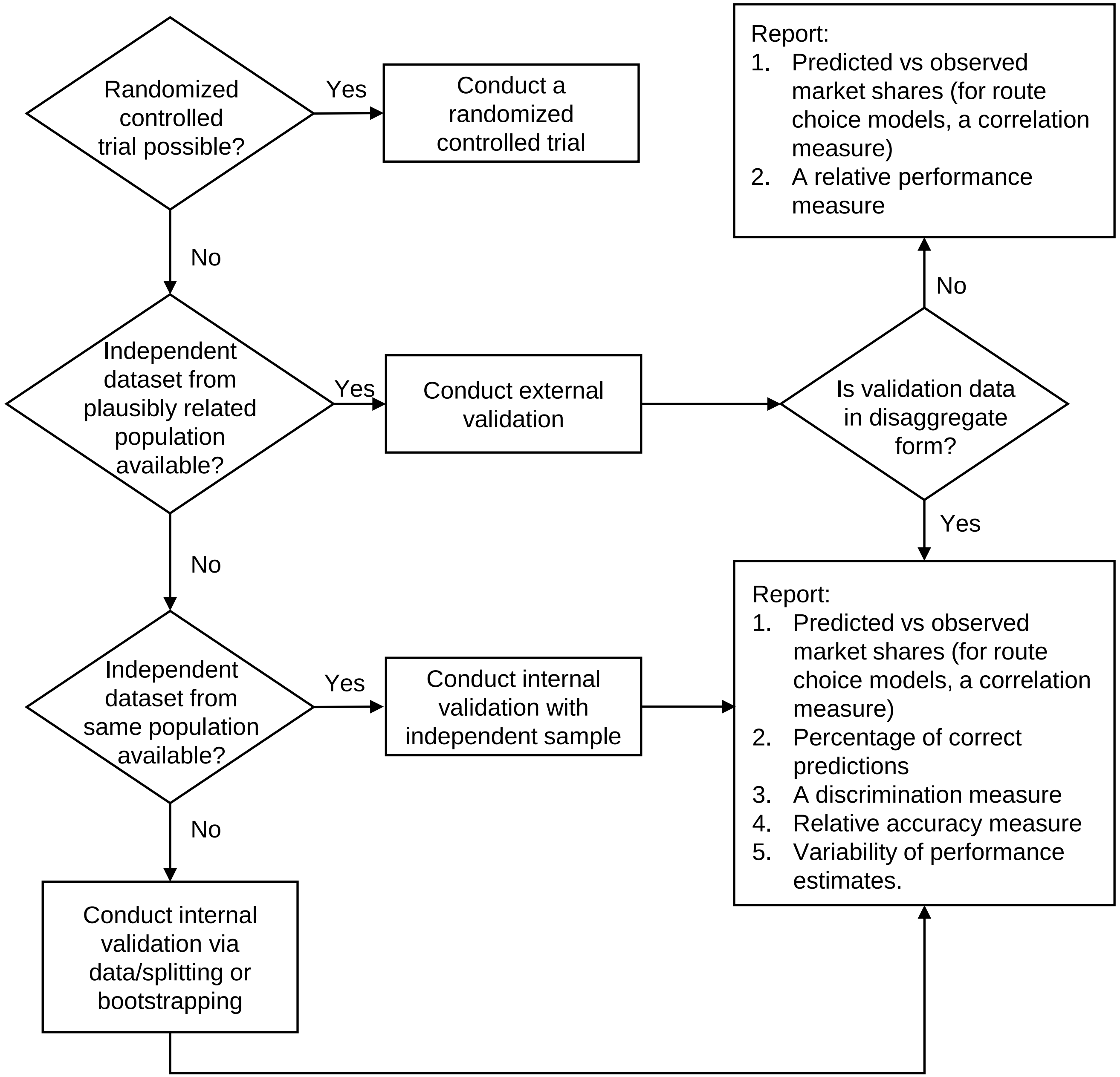
19% reported a validation measure

Method	Abbvr.	Percentage
Holdout validation	HOV	58.8%
Repeated learning-testing cross validation	RLT	26.5%
Validation against an independent sample	IS	8.8%
K-fold cross validation	K-CV	2.9%
Other sample splitting methods	SS-O	2.9%

Validation and reporting practices in the transportation academic literature

Performance measure	Abbrv.	Percentage
Log-likelihood	LL	32.6%
Percentage of correct predictions or First Preference Recovery	FPR	23.3%
Predicted vs observed market outcomes	PVO	16.3%
Mean absolute error	MAE	14.0%
Mean log-likelihood loss	MLLL	14.0%
Root mean square error	RMSE	9.3%
Absolute percentage error	APE	4.7%
Mean absolute percentage error	MAPE	4.7%
Chi-square	CHISQ	4.7%
Rho-Square	RHOSQ	4.7%
Transfer index	TI	4.7%
% clearly right (t)	%CR	2.3%
Sum of square error	SSE	2.3%
Brier Score	BS	2.3%
Concordance index	C	2.3%
Correlation	Corr	2.3%
Transferability test statistic	TTS	2.3%
All others		0.0%

Validation and reporting practices in the transportation academic literature



Recommended validation practices given available resources

Towards better validation practices in the field

- **Make model validation mandatory:**
  - Non-negotiable part of model reporting and peer-review in academic journals for any study that provides policy recommendations.
  - Cross-validation is the norm in machine learning studies.

- **Share benchmark datasets:**
  - A fundamental limitation in the field is the lack of benchmark datasets and a general culture of sharing code and data.

- **Incentivize validation studies:**
  - Lot of emphasis on theoretically innovative models.
  - Encourage submissions that focus on proper validation of existing models and theories.

- **Draw and enforce clear reporting guidelines:**
  - In addition to detailed information of survey characteristics such as sampling method, discussion on representativeness of the data, validation reporting is required.
  - Efforts to improve reporting are well documented in other fields (i.e. STROBE statement (von Elm et al., 2007))

Wait a minute…

"I'm not validating my model because I'm not trying to build a predictive framework. I'm trying to learn about travel behavior"

The more orthodox the type of analysis conducted (such as the dimensions of travel behavior covered in this study), **the stronger the onus of validation.**

Wait a minute…

"Does every study that uses a discrete choice model should be conducting validation?"

In short, yes. At the very least, **any article that makes policy recommendations should be subject to proper validation** given the dependence of the field on cross-section observational studies, and the lack of a feedback loop in academia.

Wait a minute…

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

There is a myriad of reasons why some **skepticism is warranted** against any particular model outcome. the most obvious one being model overfitting.

Finally

Better validation practices will not solve the credibility crisis in the field, but it's a step in the right direction.

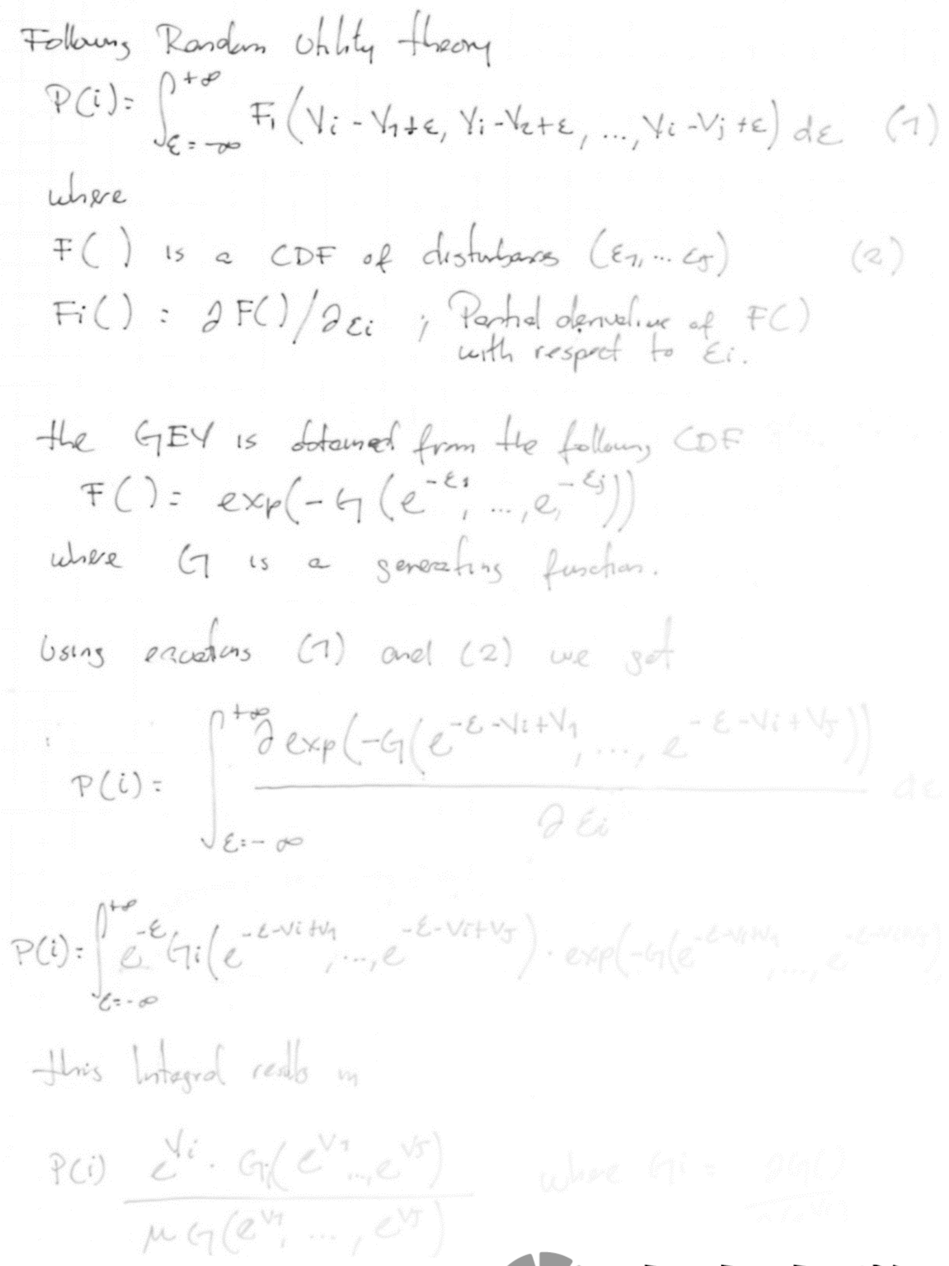
Model validation is **no solution to the causality problem** in the field, but we want to underscore that **the reliance on observational studies inherent to the field demands more stringent controls to improve external validity of results**.

References:

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

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

3. Parady G., Ory, D., Walker, J. (2019) “The overreliance on statistical goodness of fit and under-reliance on empirical validation in discrete choice models: A review of validation practices in the transportation academic literature” Presented at the 6th International choice modelling conference, Kobe, Japan, August 19-21, 2019.



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

Type	Measure	Abbrv.	Equation	Notes
Direct predictive accuracy measure	Predicted vs observed outcomes	PVO	-	Simple comparison of predicted and observed outcomes (i.e. market shares, trips by mode, etc.). Usually in the form of a table or plot. No prediction accuracy statistics are calculated.
	Percentage of correct predictions or First Preference Recovery	FPR	$\frac{100}{N}\sum_{n=1}^{N} \hat{y}_n^c = y_n^c$	y_n^c is the observed choice made by individual n, and $\hat{y}_n^c$ is the choice with the highest predicted probability.
	% clearly right (t)	%CR	$\frac{100}{\mathrm{N}}\sum_{n=1}^{N} \hat{P}(y_n^c) > t$	$\hat{P}(y_n^c)$ is the estimated choice probability of the chosen alternative. $\hat{P}(y_n^{!c})$ is the estimated choice probability of an alternative other than the chosen one.
	% clearly wrong (t)	%CW	$\frac{100}{\mathrm{N}}\sum_{n=1}^{N} \hat{P}(y_n^{!c}) > t$	
	% unclear (t)	%U	100 – (%$_{\text{clearly right}}$ (t) + %$_{\text{clearly wrong}}$ (t))	
	Fitting factor	FF	$\frac{1}{\mathrm{N}}\sum_{n=1}^{N} \hat{P}(y_n^c)$	$\hat{P}(y_n^c)$ is the estimated choice probability of the chosen alternative.
	Correlation	Corr	$\mathrm{corr}(s, \hat{s})$	Correlation between predicted and observed outcomes. s is a continuous aggregate outcome measure (i.e. train ridership, etc.)
	Brier Score	BS	$\frac{1}{\mathrm{N}}\sum_{n=1}^{N}\sum_{m=1}^{M}\left(\hat{P}(y_{nm}) - y_{nm}\right)^2$	$\hat{P}(y_{nm})$ is the predicted probability that individual n chooses alternative m and y_{nm} is the actual outcome variable valued 0 or 1. In the particular case of a binary choice, the second summation sign disappears.
	$\rho^2, \bar{\rho}^2$	RHOSQ	$\rho^2 = 1 - \frac{LL(\hat{\beta})}{LL(0)};\ \bar{\rho}^2 = 1 - \frac{AIC}{LL(0)}$	LL(0) is the log-likelihood when all parameters are zero. LL(β) is the maximized likelihood. K is the number of freely estimated parameters in the model.

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

Type	Measure	Abbrv.	Equation	Notes
Relative predictive accuracy measures	Absolute percentage error	APE	$100 \cdot \left\lvert \frac{\hat{s}_m - s_m}{s_m} \right\rvert$	M is the number of alternatives in the choice set. s_m is an aggregate outcome measure, such as the market share of alternative m (i.e. modal market share), choice frequency, etc. $\hat{P}(y_{nm})$ is the predicted probability that individual n chooses alternative m and y_{nm} is the actual outcome variable valued 0 or 1. In the particular case of a binary choice, the second summation sign disappears.
	Sum of square error	SSE	$\sum_{m=1}^{M} (\hat{s}_m - s_m)^2$	
	Root sum of square error	RSSE	$\sqrt{\sum_{m=1}^{M} (\hat{s}_m - s_m)^2}$	
	Mean absolute error	MAE	$\frac{1}{M} \sum_{m=1}^{M} \lvert \hat{s}_m - s_m \rvert$	
	Mean absolute percentage error	MAPE	$\frac{100}{M} \sum_{m=1}^{M} \left\lvert \frac{\hat{s}_m - s_m}{s_m} \right\rvert$	
	Mean squared error	MSE	$\frac{1}{M} \sum_{m=1}^{M} (\hat{s}_m - s_m)^2$	
	Root mean square error	RMSE	$\sqrt{\frac{1}{M} \sum_{m=1}^{M} (\hat{s}_m - s_m)^2}$	
	χ^2 test	CHISQ*	$\sum_{m=1}^{M} \frac{(f_m - E(f_m))^2}{E(f_m)}$	f_m is the observed choice frequency of alternative m, and $E(f_m)$ is the expected choice frequency

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

Type	Measure	Abbrv.	Equation	Notes
Relative predictive accuracy measures	Maximum market share deviation	MSD	$\max S$	S is the set of all market share deviations
	Log-likelihood	LL	$\sum_{n=1}^{N}\sum_{m=1}^{M} c_{nm} log \hat{P}(y_n)$	c_{nm} is a variable that takes value 1 if alternative m was chosen by individual n, and 0 otherwise.
	Mean log-likelihood loss	MLLL	$\frac{1}{\mathrm{R}}\sum_{r} -\frac{1}{VS_r} LL$	Where LL is the log-likelihood, VS is the size of the validation (holdout) sample r, and R is number of validation samples generated.
	likelihood ratio test (LR), AIC	f(LL)	$AIC = LL(\hat{\boldsymbol{\beta}}) - K$ $LR = -2\left(LL(0) - LL(\hat{\boldsymbol{\beta}})\right)$	LL(0) is the log-likelihood when all parameters are zero. LL(β) is the maximized likelihood. K is the number of freely estimated parameters in the model.