Decision-aid methodologies in transportation

CIVIL-557

Modelling transportation systems
2. Trip generation models

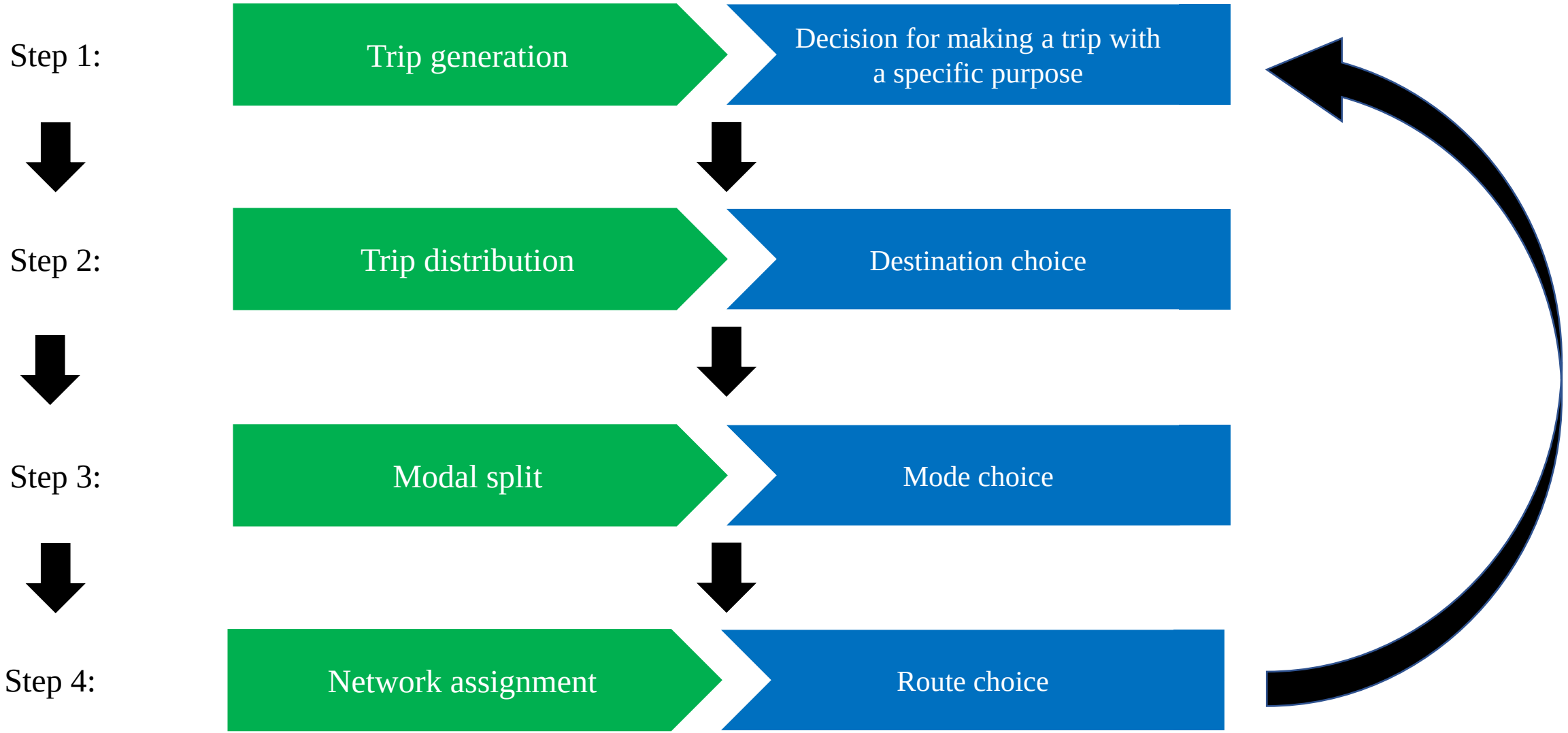
Evangelos Paschalidis

TRANSP-OR EPFL

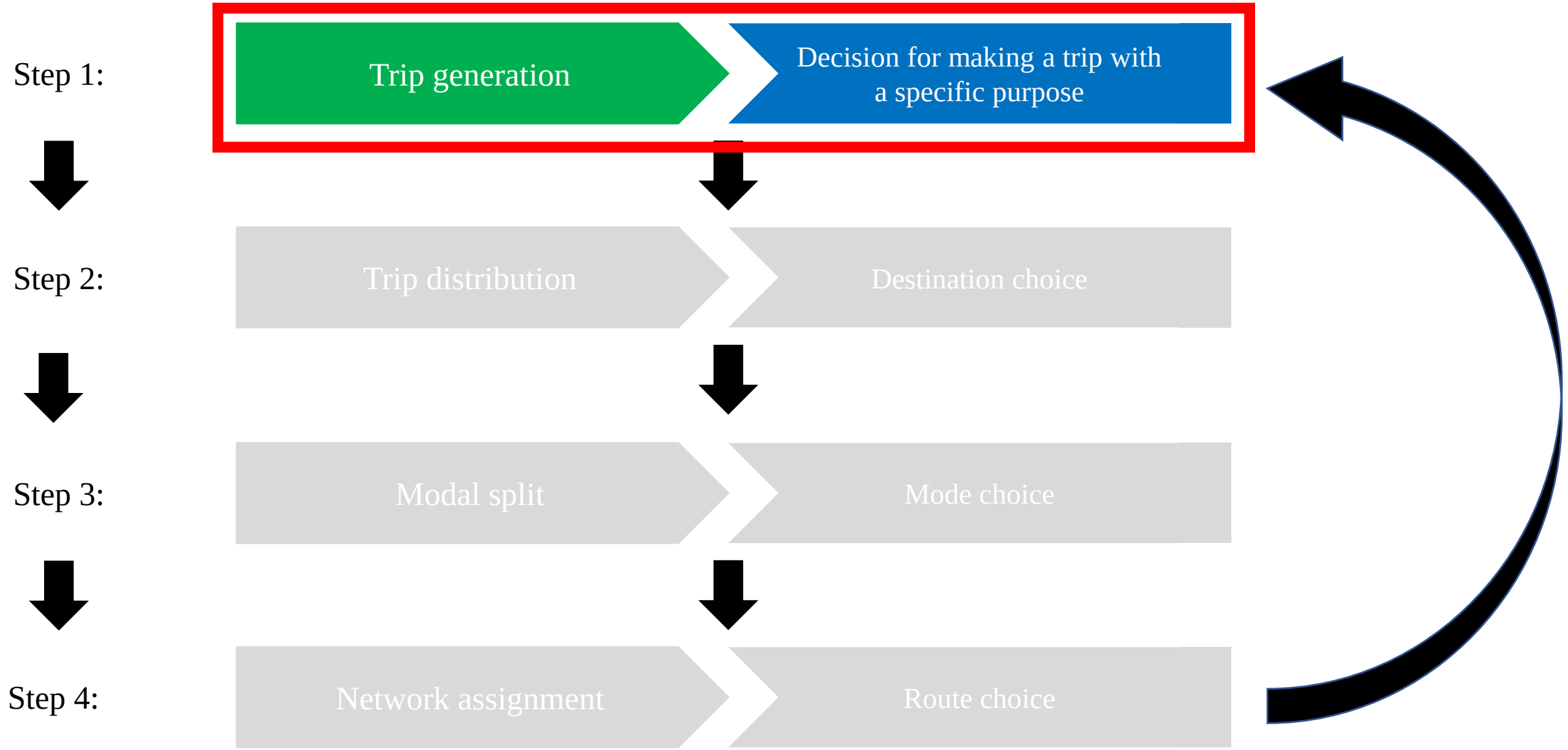
Previously…

- The definition of transportation modelling
- The purpose of transportation modelling
- Terminology
- Model specification, calibration, and validation
- The 4–step model
- Data collection – Sampling

The 4-step model



The 4-step model



Today

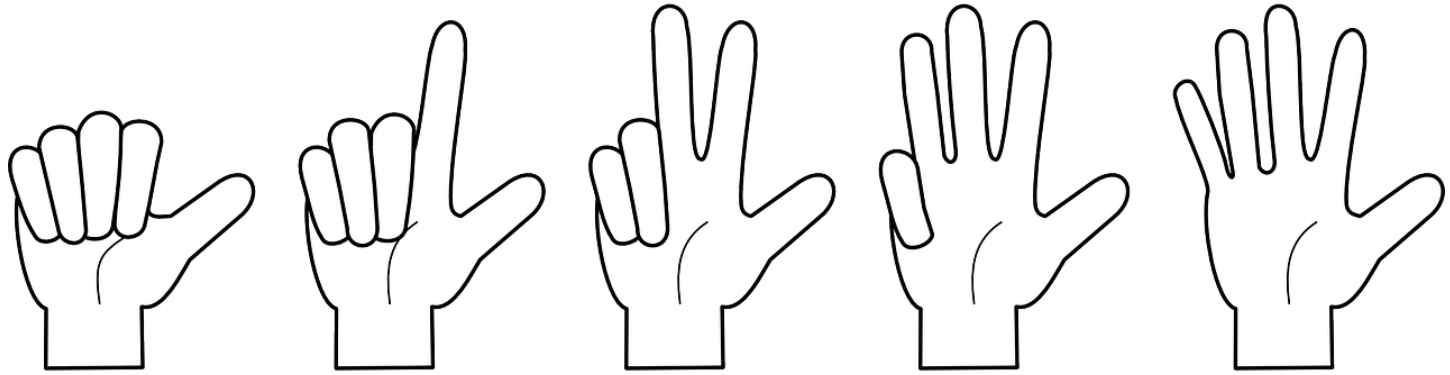
Trip generation modelling (step 1 of the 4–step model)

- Aim – motivation – purpose
- Terminology
- Models
  - Cross classification – category analysis
  - Growth factor models
  - Linear regression

Trip generation

Aim - Motivation

- Compute the number of trips associated with each type of activity (work, leisure, education etc.)
  - Model the trips generated by and attracted to each zone of the study area
  - The trips are typically modelled at the household and zonal level
  - Trip generation modelling does not deal with how demand is split between modes or destinations



Trip generation: A clarification

Trip generation modelling ≠ “Generated trips” under some scenario

- Increase of vehicles or passengers in a location may be a result of:
  - changing routes (Step 4 - assignment model)…
  - … or people may change destination (Step 2 – trip distribution model)…
  - … or traffic may increase because people travel more with car (Step 3 – mode split model)

Terminology (useful reminder)

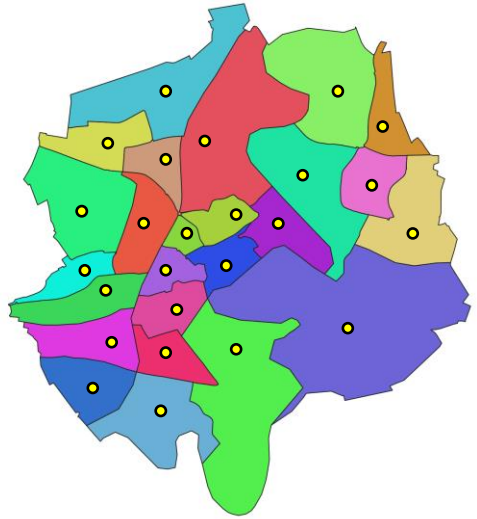
- **Trip or Journey:** A one-way movement from a point of origin to a point of destination. We are usually interested only in vehicular trips (even though we shouldn't!!).
- Trip ends: total number of trips generated in a zone
- We estimate the trip ends via ***trip rates***
- Trip rate: typical number of trips (per unit) made by
  - Different types of people
  - In different types of areas
  - In different types of buildings
  - For different purposes

Rule of thumb: Trip rates remain quite stable over time, they are our starting point in the modelling process

Zone (another useful reminder)

Traffic zone

- The unit of transport analysis
- Typically homogeneous
  - Land use
  - Transport network
  - Area of main centres of activities
  - Boundaries of administrative units
- Centroid: a point in each zone that is used to define the centre of activities in a zone and link it to the traffic network (all demand is produced or attracted at the centroid of each zone)

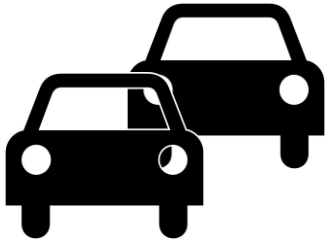


Characterisation of trips

- Purpose: work, education, shopping, leisure, escort trips, other…
  - Work and education are typically called mandatory trips
- Time of the day: Peak hours (morning & afternoon peak), off-peak hours
- Individual characteristics: income, car ownership, household size and structure, etc.

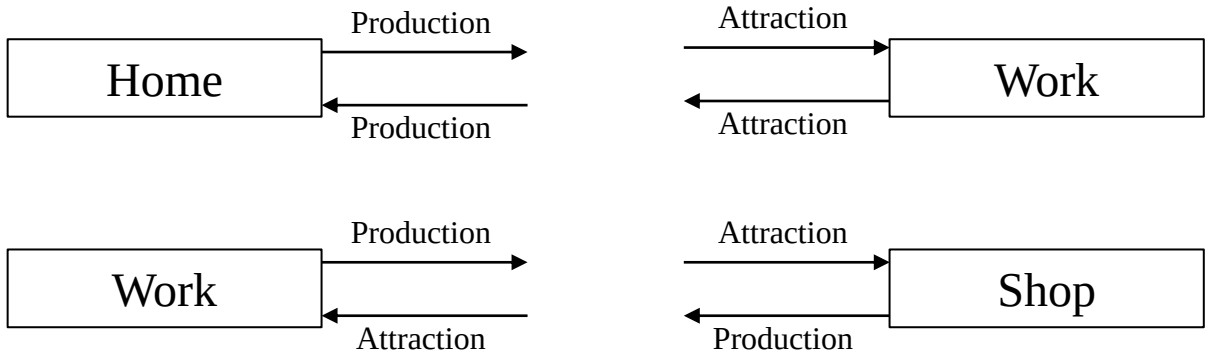






Production and attraction

- Two different ways for the trip generation process: production and attraction
- Two main types of trips:
  - Home-based (HB) Trip: Home of the trip maker is either the origin or the destination of the journey
  - Non-home-based (NHB) Trip: Neither end of the trip is the home of the traveller.
- Trip Production: The home end of an HB trip or as the origin of an NHB (all trips leaving or arriving home are ***'produced'*** at the home location)
- Trip Attraction: The non-home end of an HB trip or the destination of an NHB trip
- Trip Generation: The total number of trips generated by households in a zone, both HB or NHB.

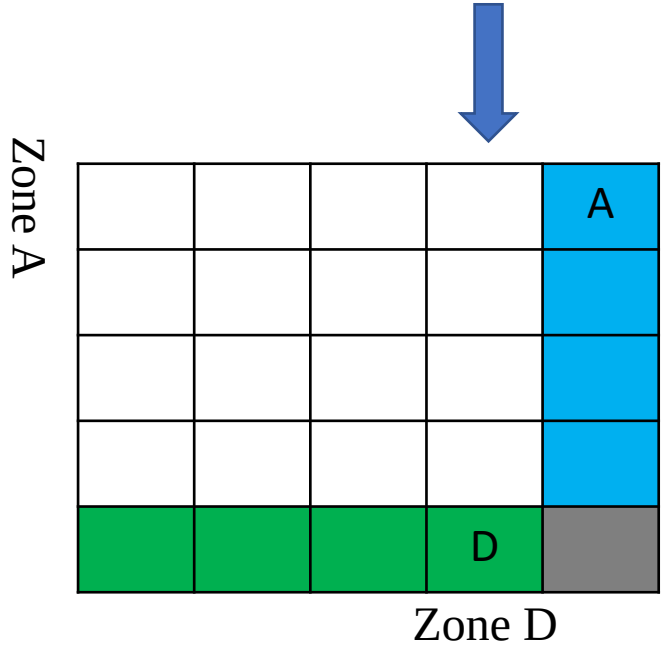


Production and attraction

- Each trip is produced in a zone and attracted to a zone
- At this stage we do not examine the exact origin-destination pair (this is done in Step 2)
- NHB trips are more difficult to model. Simplification e.g. 30% of trips attracted to a zone

Number of trips from/to zone B, as long as they are attracted to zone B

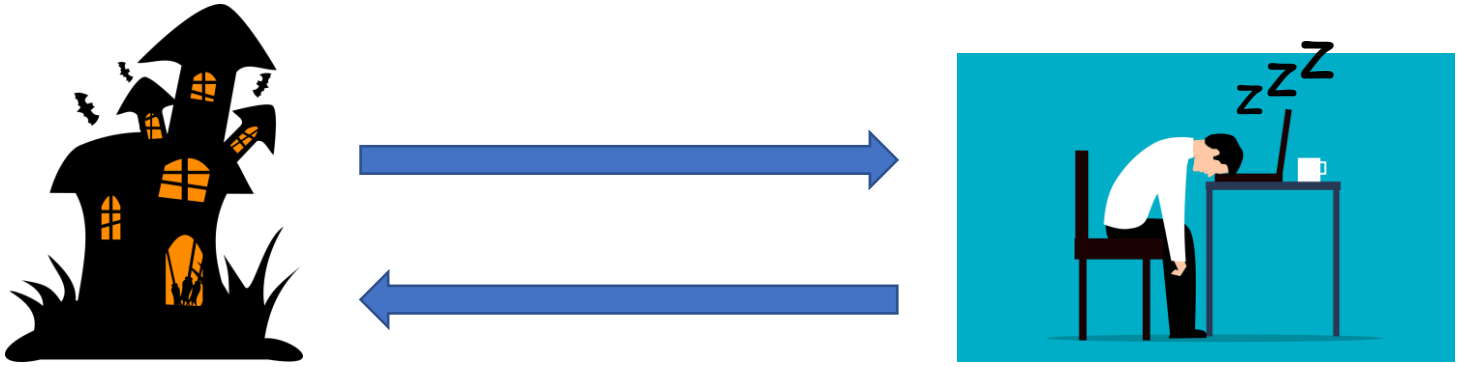
Number of trips from/to zone A, as long as they are produced in zone A



Production and attraction

Small example: I live in zone A and work in zone B

- Origin-destination format: (A, B) trip to work, and (B, A) return trip
- Production-attraction format: Two trips produced by zone A and two trips attracted to zone B

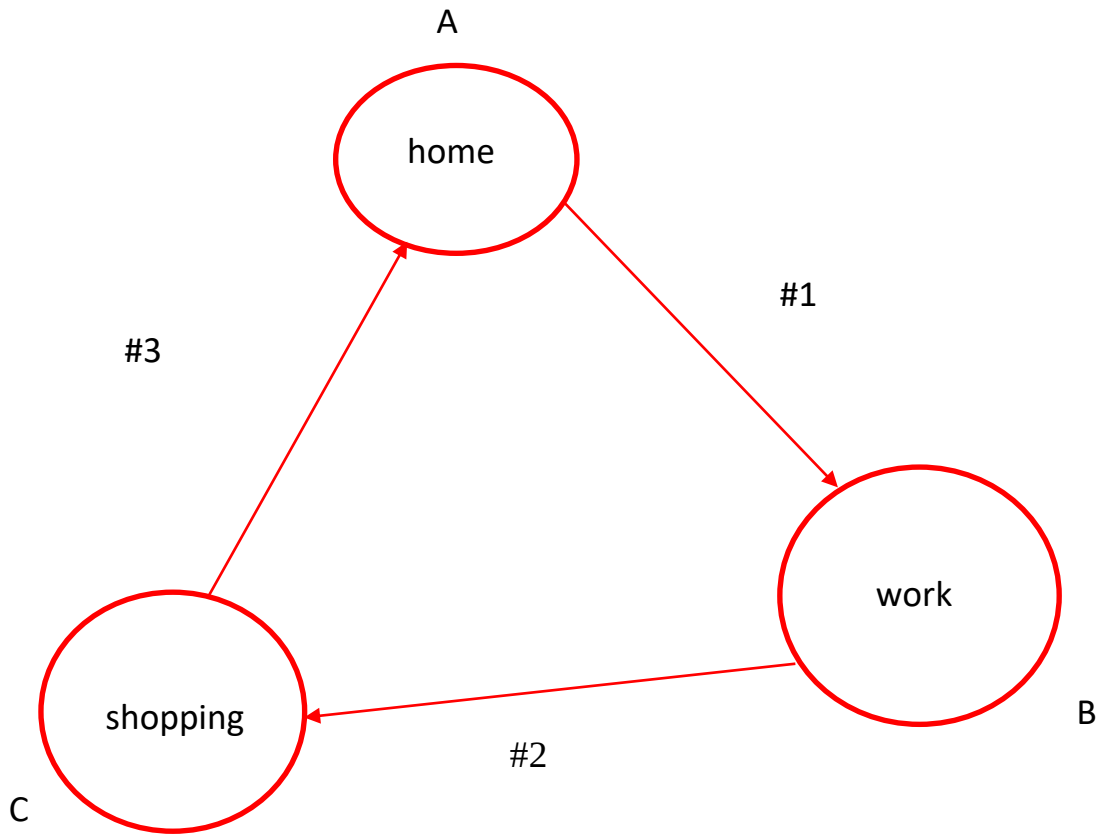


Zone A: Home

Zone B: Work

Production and attraction: Example

Following the direction of the arrows, define the production-attraction and origin-destination points for each of the trips:



Production and attraction: Example

Following the direction of the arrows, define the production-attraction and origin-destination points for each of the trips:

Trip #1:

Origin A – Destination B

Production A – Attraction B

Trip #2:

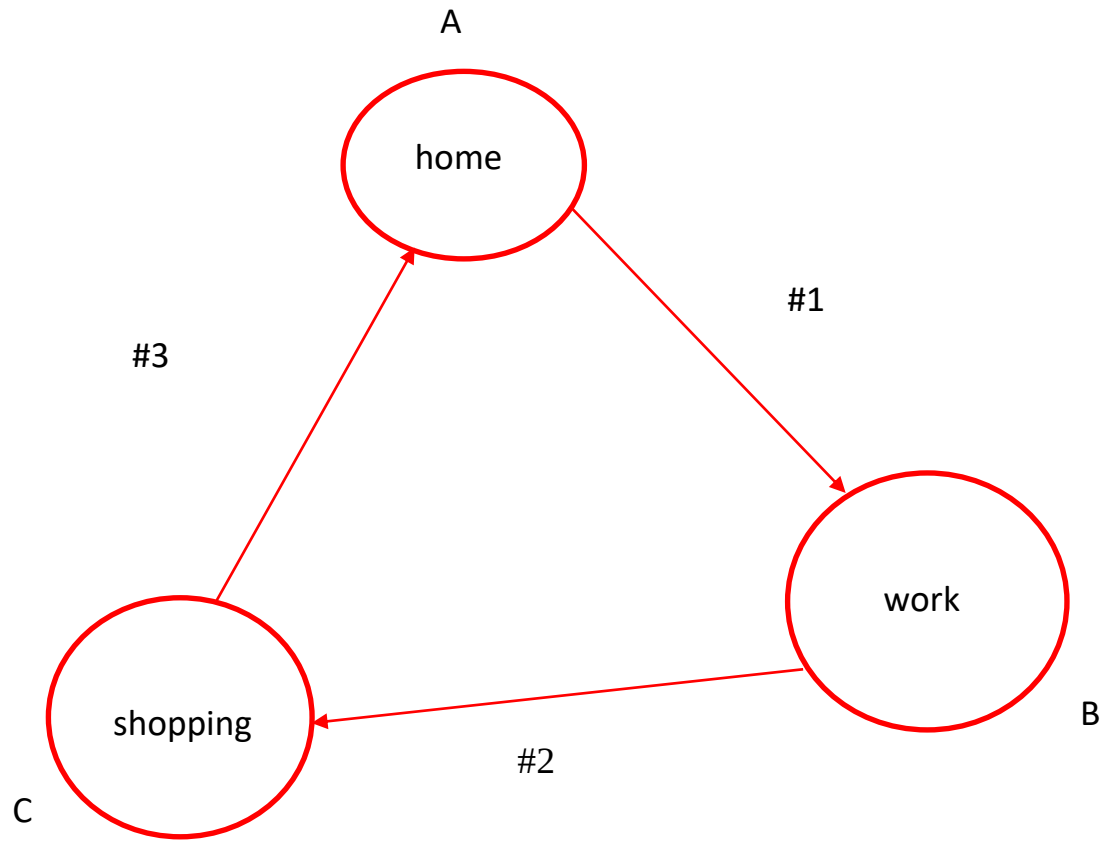
Origin B – Destination C

Production B – Attraction C

Trip #3:

Origin C – Destination A

Production A – Attraction C



Trip generation models categories

- Models of production
- Models of attraction
- We use different models for production and attraction
- Different models for trips outside the study area or freight transport
- Our Objective: Estimate how many trips are generated and attracted by each zone (we need to model)
- In a closed system: production trips = attraction trips

Factors affecting trip generation

Personal trip productions

- Socioeconomic factors: as specified earlier (HH size, income etc.)
- Land use: value of land, residential density
- Accessibility e.g. availability of public transport in urban areas

Personal trip attractions

- Roofed space available for industrial, commercial and other services
- Zonal employment – Number of employment positions

Trip rates example

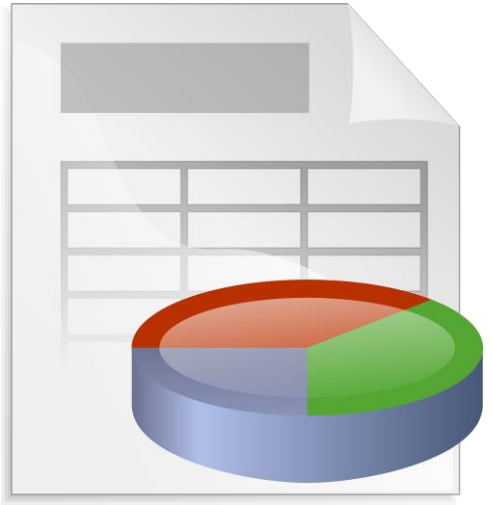
- Trip ends typically have quite detailed segmentation
  - You can also check last week's demand segmentation
- Example:
  - Trip purpose: commuting, leisure, shopping, …
  - Household income: Low, average, high
  - Number of cars: 0, 1, 2+ (0, 1+ for example simplification)
  - Number of children, 0, 1, 2, 3+
  - Area of living: rural, urban

Purpose	Income	Cars	No of children	Area
Commute	Low	0	0	Urban
Commute	Low	0	0	Suburbs
Commute	Low	0	1	Urban
Commute	Low	0	1	Suburbs
Commute	Low	0	2	Urban
Commute	Low	0	2	Suburbs
Commute	Low	0	3+	Urban
Commute	Low	0	3+	Suburbs
Commute	Low	1+	0	Urban
Commute	Low	1+	0	Suburbs
Commute	Low	1+	1	Urban
Commute	Low	1+	1	Suburbs
Commute	Low	1+	2	Urban
Commute	Low	1+	2	Suburbs
Commute	Low	1+	3+	Urban
Commute	Low	1+	3+	Suburbs
⋮				
Leisure	High	1+	0	Urban
Leisure	High	1+	0	Suburbs
Leisure	High	1+	1	Urban
Leisure	High	1+	1	Suburbs
Leisure	High	1+	2	Urban
Leisure	High	1+	2	Suburbs
Leisure	High	1+	3+	Urban
Leisure	High	1+	3+	Suburbs

Information required

We need information (data) about the:

- Number of households of each type (based on the grouping we decided before)
- Built space for different types of land use (shops, offices, hotels, restaurants etc…)
  - Some trips may be expressed as number of trips per unit of built space
- If we already have information on current trip rates: we know the base year scenario
- We still need a model for a future scenario e.g. what happens if we build new houses or new shops in a zone.



Modelling process – Summary

1. Grouping of decision making units (e.g. household type)
2. Aggregation in time periods (instead of individual level trips)
3. Segregation per trip purpose (mainly work, leisure, and shopping)

Trip generation models

Deterministic models

- Cross classification – category analysis
- Growth factor models

Stochastic models

- Linear regression

Category analysis

- Estimate of the variable of interest (e.g trip productions per household – HH) as a function of the HH characteristics
- Households are classified based on their characteristics
- Trip generation rates are computed from data about the current condition (base scenario)
- Future trip generation rates are computed based on a scenario
- Assumption: Trip generation rates and stable over time and HH characteristics

Considerations:

- At what dimension (how many levels) to cross-classify e.g. HH size, car ownership, income etc…
- Choice of a category that is relatively stable over time

Category analysis

The category analysis is based on the computation of trip rates (trips per HH):

$$t^p(h) = T^p(h)/H(h)$$

h: household of type *h*

p: trip purpose

$t^p(h)$: average number of trips per household, with purpose *p*

$T^p(h)$: observed (total) trips per *h* and *p*

H(h): number of households of type *h*

Example 1

Data

- The (average) number of trips made by HH
- The number of cars owned
- Distribution of HHs by car ownership, HH size, etc.

Question

- What is the average number of trips per HH?

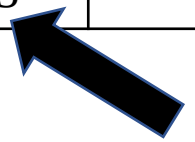
Example 1

Col. 1	Col. 2	Col. 3	Col. 4	Col. 5	Col. 6
Car Ownership	Av. number of trips per HH	% HH			
0	6	34%			
1	6.78	47%			
2+	7.52	19%			

Example 1

Average number of trips by car ownership = trips made by HH * % HH of the population

Col. 1	Col. 2	Col. 3	Col. 4	Col. 5	Col. 6
Car Ownership	Av. number of trips per HH	% HH	Col. 2 * Col. 3		
0	6	34%	2.04		
1	6.78	47%	3.1866		
2+	7.52	19%	1.4288		
			6.655		

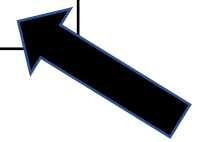


Average of population

Example 1

Scenario: we change the distribution of HH

Col. 1	Col. 2	Col. 3	Col. 4	Col. 5	Col. 6
Car Ownership	Av. number of trips per HH	% HH	Col. 2 * Col. 3	New % of HH (scenario)	
0	6	34%	2.04	15%	0.9
1	6.78	47%	3.1866	55%	3.729
2+	7.52	19%	1.4288	30%	2.256
			6.655		6.885



New average

Example 1

Scenario: we change the distribution of HH

Col. 1	Col. 2	Col. 3	Col. 4	Col. 5	Col. 6	Col. 7
Car Ownership	Av. number of trips per HH	% HH	Col. 2 * Col. 3	New % of HH	Col. 2 * Col. 5	Difference
0	6	34%	2.04	15%	0.9	-1.14
1	6.78	47%	3.1866	55%	3.729	+0.5424
2+	7.52	19%	1.4288	30%	2.256	+0.8272
			6.655		6.885	+0.2296

Difference

Limitations of Example 1

- We considered only one category to group the households
  - Maybe we ignored some important household characteristics by generalising too much?
  - Could this have an impact on the future scenario?
- Results of the future scenario are significantly affected by number of cars
  - What if we overestimated trip rates due to this simplification?
- Solution: Let’s add another category to further segment the households

Example 2

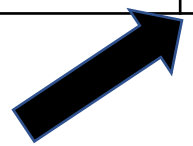
Col. 1	Col. 2	Col. 3	Col. 4	Col. 5	Col. 6	Col. 7	Col. 8
Car Ownership	HH Size	Av. trips per HH	%HH				
0	1	4.25	4%				
0	2	5.666667	15%	← 34%			
0	3+	6.8	15%				
1	1	5	7%				
1	2	6.222222	18%	← 47%			
1	3+	7.818182	22%				
2+	1	5.6	5%				
2+	2	7.2	5%	← 19%			
2+	3+	8.777778	9%				

We add one more group (HH size)

Example 2

Col. 1	Col. 2	Col. 3	Col. 4	Col. 5	Col. 6	Col. 7	Col. 8
Car Ownership	HH Size	Av. trips per HH	%HH	(Col. 3)* (Col. 4)			
0	1	4.25	4%	0.17			
0	2	5.666667	15%	0.85			
0	3+	6.8	15%	1.02			
1	1	5	7%	0.35			
1	2	6.222222	18%	1.12			
1	3+	7.818182	22%	1.72			
2+	1	5.6	5%	0.28			
2+	2	7.2	5%	0.36			
2+	3+	8.777778	9%	0.79			
				6.65			

Average trips remain the same for the base scenario



Example 2

Col. 1	Col. 2	Col. 3	Col. 4	Col. 5	Col. 6	Col. 7	Col. 8
Car Ownership	HH Size	Av. trips per HH	%HH	(Col. 3)* (Col. 4)	New % HH		
0	1	4.25	4%	0.17	5%		
0	2	5.666667	15%	0.85	5%		
0	3+	6.8	15%	1.02	5%		
1	1	5	7%	0.35	10%		
1	2	6.222222	18%	1.12	20%		
1	3+	7.818182	22%	1.72	25%		
2+	1	5.6	5%	0.28	10%		
2+	2	7.2	5%	0.36	10%		
2+	3+	8.777778	9%	0.79	10%		
				6.65			

Example 2

Col. 1	Col. 2	Col. 3	Col. 4	Col. 5	Col. 6	Col. 7	Col. 8
Car Ownership	HH Size	Av. trips per HH	%HH	(Col. 3)* (Col. 4)	New % HH	(Col. 6)* (Col. 3)	(Col. 7)-(Col. 5)
0	1	4.25	4%	0.17	5%	0.2125	+0.0425
0	2	5.666667	15%	0.85	5%	0.283333	-0.56667
0	3+	6.8	15%	1.02	5%	0.34	-0.68
1	1	5	7%	0.35	10%	0.5	+0.15
1	2	6.222222	18%	1.12	20%	1.244444	+0.124444
1	3+	7.818182	22%	1.72	25%	1.954546	+0.234545
2+	1	5.6	5%	0.28	10%	0.56	+0.28
2+	2	7.2	5%	0.36	10%	0.72	+0.36
2+	3+	8.777778	9%	0.79	10%	0.877778	+0.087778
				6.65		6.692601	+0.032601

Discussion of Example 2

- Average trips per household remain the same for the base scenario
- In example 2, forecast is 6.69 trips per household, as opposed to 6.88 in example 1, why?
  - Greater degree of classification (as in example 2) allows greater sophistication in assumptions and hence possibility of more realistic representation of real life.
  - If we generalise too much, we may overestimate (or underestimate) the trip rates
  - The difference between the two (synthetic) examples is small but much larger differences can be observed in reality

Modelling car ownership

- Aim: To model the number of cars owned by households, where HH are categorised by location (URBAN/RURAL), and HH size (number of adult members)
- A survey was carried out at two locations (Urban and Rural) and data presented as follows:

Col. 1	Col. 2	Col. 3	Col. 4	Col. 5	Col. 6
Location	HH Size	Cars owned by HH	No of HH		
URBAN	1	0000000011	10		
URBAN	2	000000111122	12		
URBAN	3+	0122	4		
RURAL	1	00111	5		
RURAL	2	00111122	8		
RURAL	3+	01222	5		

(a) The number of digits indicate the number of HH (e.g. the ten digits in 0000000011 = 10 HH) (b) The value of each digit indicates the number of cars owned by a HH (e.g. ‘2’ means two cars and ‘0’ means no cars).

Modelling car ownership

Col. 1	Col. 2	Col. 3	Col. 4	Col. 5	Col. 6
Location	HH Size	Cars owned by HH	No of HH	% of HH	
URBAN	1	0000000011	10	22.7%	
URBAN	2	000000111122	12	27.3%	
URBAN	3+	0122	4	9.1%	
RURAL	1	00111	5	11.4%	
RURAL	2	00111122	8	18.1%	
RURAL	3+	01222	5	11.4%	
			44		

Modelling car ownership

Col. 1	Col. 2	Col. 3	Col. 4	Col. 5	Col. 6
Location	HH Size	Cars owned by HH	No of HH	% of HH	No of cars
URBAN	1	0000000011	10	22.7%	2
URBAN	2	000000111122	12	27.3%	8
URBAN	3+	0122	4	9.1%	5
RURAL	1	00111	5	11.4%	3
RURAL	2	00111122	8	18.1%	8
RURAL	3+	01222	5	11.4%	7
			44		

Modelling car ownership

Col. 1	Col. 2	Col. 3	Col. 4	Col. 5	Col. 6	Co. 7
Location	HH Size	Cars owned by HH	No of HH	% of HH	No of cars	Avg No of Cars per HH
URBAN	1	0000000011	10	22.7%	2	0.20 (=2/10)
URBAN	2	000000111122	12	27.3%	8	0.67 (=8/12)
URBAN	3+	0122	4	9.1%	5	1.25
RURAL	1	00111	5	11.4%	3	0.60
RURAL	2	00111122	8	18.1%	8	1.00
RURAL	3+	01222	5	11.4%	7	1.40
			44			

Modelling car ownership

Example 3b: Using the data from 3a predict the average number of cars per household for a similar population distributed as:

	URBAN 1	URBAN 2	URBAN 3+	RURAL 1	RURAL 2	RURAL 3+
Case 1	10%	20%	20%	10%	20%	20%
Case 2	5%	10%	10%	15%	30%	30%

Modelling car ownership

3b-1

	URBAN 1	URBAN 2	URBAN 3+	RURAL 1	RURAL 2	RURAL 3+
Case 1	10%	20%	20%	10%	20%	20%

Col. 1	Col. 2	Col. 3	Col. 4	Co. 5
Location	HH Size	% of HH	No of cars	Col. 3 * Col. 4
URBAN	1	10%	0.20	0.020
URBAN	2	20%	0.67	0.134
URBAN	3+	20%	1.25	0.25
RURAL	1	10%	0.60	0.06
RURAL	2	20%	1.00	0.200
RURAL	3+	20%	1.40	0.280
				0.944

Modelling car ownership

3b-2

	URBAN 1	URBAN 2	URBAN 3+	RURAL 1	RURAL 2	RURAL 3+
Case 2	5%	10%	10%	15%	30%	30%

Col. 1	Col. 2	Col. 3	Col. 4	Co. 5
Location	HH Size	% of HH	No of cars	Col. 3 * Col. 4
URBAN	1	5%	0.20	0.010
URBAN	2	10%	0.67	0.067
URBAN	3+	10%	1.25	0.125
RURAL	1	15%	0.60	0.090
RURAL	2	30%	1.00	0.300
RURAL	3+	30%	1.40	0.420
				1.012

Comments on category analysis

Advantages

- No prior assumptions about the shapes of relationships are required, i.e. no formula needed
- Analysis can be carried out on disaggregated data
- Independence between HH characteristics variables and zones structure
- Independence between the different HH characteristics

Disadvantages

- Unless using regression analysis to select variables and levels of variables, there are no goodness of fit tests
- Extrapolation not possible; using open ended (continuous) variables levels can have difficulties
- Large samples required: preferably 20 to 50 observations per cell

Demand growth scenarios

- Demand growth scenarios: we assume different trip generation scenarios, following a change in our system
- Test the project on all the scenarios, some will never happen
- Not a robust scientific method on how to develop demand growth scenarios. Considerations:
  - Trip rates: Standard trip rates are the average of observed data. We also consider different values, especially if we have knowledge of the local area
  - Employment growth: Employment rates, sectors, wider economic impacts, also in the surrounding zones
  - Trends: Working from home, ride sharing, autonomous vehicles (future scenario)
  - Population growth: Different assumptions, people moving in and out, new buildings
  - Shopping and leisure: Competition between retail and leisure centres
  - Mode-specific factors: Car ownership, investment on walking or cycling infrastructure, capacity of public transport, parking policy

Expansion (or Growth) Factor

Main formula

$$T_i = F_i \, t_i$$

where T_i and t_i are respectively future (scenario) and current (base year) trip origins in zone i, and F_i is a growth factor

Normally the growth factor is related to variables such as population (P), income (I) and car ownership (C), in a function such as:

$$F_i = \frac{f\left(P_i^d, I_i^d, C_i^d\right)}{f\left(P_i^c, I_i^c, C_i^c\right)}$$

where f can even be a direct multiplicative function with no parameters, and the superscripts d and c denote the design and current years respectively

Similarly, this method can be applied to trips destinations

Expansion (or Growth) Factor - Example

a. Base year

250 HH with a car

6 trips/day (with a car)

250 HH without a car

2.5 trips/day (w/o a car)

total trips: $t_i = 250 \times 6 + 250 \times 2.5 = 2125$ trips/day

b. Target year

500 HH with a car

0 HH without a car

How many trips/day in the target year?

Expansion (or Growth) Factor - Example

Solution

All but car ownership remain constant:

$$F_i = C_i^d / C_i^c = 1/0.5 = 2$$

$T_i = 2 \times 2125 = 4250$ trips/day

Initial formula (previous slide)

T_i=500*6=3000

The growth factor approach can be very crude – estimated (4250-3000)/3000 = 42% more trips

Expansion (or Growth) Factor

- Very simple
- Requires good base year data for all zones to obtain reliable expansion factors
- Separate expansion factors could be used for different zones or types (in inner area/outer area)
- Not sensitive to policy changes without taking into account all relevant factors
  - E.g. Based on some employment level trends, I expect a 10% population increase, hence $F_i = 1.1$
  - What if some new congestion charge is introduced?
- Growth factor methods are mostly used predict the future number of *external* trips to an area
  - usually they are not too many (so errors cannot be too large)
  - there are no simple ways to predict them

Linear regression

Statistical technique to "explain" movements (changes) in a variable (dependent variable) as a function of other variables (independent variables):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

where

Y_i : Dependent variable (e.g. number of trips of a household daily)

X_i: Independent variable (e.g. car availability)

β_0 & β_1: parameters to be estimated (β_0 is also known as the intercept term or simply constant)

Interpretation: if X_{1i} changes by one unit, then Y_i changes by β_1

ε_i : Independent and identically normally distributed (i.i.d.) error term

- A linear regression model can have many independent (explanatory) variables

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} \ldots + \beta_k X_{ki} + \varepsilon_i$$

Linear regression

Advantages: Easy to specify, estimate and interpret. Plenty of available software.

Linear regression (and all statistical models) require some assumptions to ensure the validity of the results.

⚠ Can be overused or misused. More advanced models must be used if assumptions are violated.

⚠ Trip generation is a ‘count’ outcome (i.e. number of trips). Linear regression is producing a continuous prediction $\hat{Y}$, e.g. 2.7 trips/hh or –0.14 trips/hh. Forecasts must be carefully checked

- Alternatives: Poisson regression, negative binomial regression or any other model suitable for a count outcome

Linear regression – Estimation procedures

So far, we have mentioned some parameters β associated with the impact of independent variables on the dependent variable…

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} \ldots + \beta_k X_{ki} + \varepsilon_i$$

… but how do we obtain their values?

Linear regression – Estimation – Ordinary Least Squares

Aim: minimise the squared error

$$\varepsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - X_i \hat{\beta}\right)^2$$

In matrix form:

$$\varepsilon' \varepsilon = \left(Y - X\hat{\beta}\right)' \left(Y - X\hat{\beta}\right) = Y'Y - 2\hat{\beta}'\mathrm{X}'Y + \hat{\beta}'\mathrm{X}'X\hat{\beta}$$

For $\frac{\partial \varepsilon' \varepsilon}{\partial \hat{\beta}} = 0$ we get

$$\hat{\beta} = (\mathrm{X}'X)^{-1}\,\mathrm{X}'\,Y$$

… these are the parameters that we want to estimate

Linear regression – Estimation – Standard errors

Standard errors: a measure of variability of the estimated parameters in the population

- Aim: the smaller the better
- How small? We investigate in the next slides

$$Var(\widehat{\beta}) = \sigma^2(\mathrm{X}'X)^{-1} \rightarrow SE(\widehat{\beta}) = \sqrt{\sigma^2(\mathrm{X}'X)^{-1}}$$

True variance unknown, instead we use the mean squared error (MSE):

$$MSE = \frac{1}{n-k}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \frac{SSE}{n-k},$$

SSE is the sum of squared errors

- n is the sample size, k is the number of the estimated parameters
- $E[MSE] = \sigma^2$

Linear regression – categorical independent variables

The interpretation of linear regression is one-unit change of an independent variable results to β change to the dependent variable.

Sometimes, the independent variable may not be continuous but categorical for instance *gender, highest education level, availability of car in a household, time of the day someone commutes (morning, noon, afternoon, evening)*.

If we use an independent variable X which is categorical, we must estimate different βs for each category.

" **Dummy variable trap!** We can estimate one less parameter than the total number of groups of the independent variable to avoid perfect multicollinearity

Linear regression – categorical independent variables

Let's consider a categorical variable D with 4 levels D_{1i}, D_{2i}, D_{3i}, D_{4i} where:

D_{ji} = 1 if condition is met for observation i, 0 otherwise.

A regression model would be:

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 D_{4i} + \beta_5 X_i + \varepsilon_i$$

In this case we keep $D4$ as a reference category and we are not estimating a parameter for it.

- E.g. $\beta_4 = 0$

Linear regression – categorical independent variables

If we attempt to estimate parameters for **all levels**:

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 D_{4i} + \beta_5 X_i + \varepsilon_i$$

By definition: $D_{1i} + D_{2i} + D_{3i} + D_{4i} = 1$

Hence:

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 (1 - D_{1i} + D_{2i} + D_{3i}) + \beta_5 X_i + \varepsilon_i$$

$$Y_i = (\beta_0 + \beta_4) + (\beta_1 - \beta_4) D_{1i} + (\beta_2 - \beta_4) D_{2i} + (\beta_3 - \beta_4) D_{3i} + \beta_5 X_i + \varepsilon_i$$

- Every group is a linear combination of the others plus the intercept – perfect multicollinearity ($X'X$ non-invertible)
- Cannot estimate the sums and differences of parameters separately Solution: Do not estimate β_4 (in general estimate one parameter less than the number of categories)

Linear regression – Estimation output

Using any linear regression software we would receive an output like this:

- Dependent variable: number of trips of a household per day

We know what are these…

What is this??

And this??

Parameter	Estimate	s.e.	t-ratio	p-value
β_0	0.883	0.047	18.655	0.000
$\beta_{(\text{Household size})}$	0.418	0.004	97.635	0.000
$\beta_{(\text{Number of vehicles})}$	0.263	0.005	53.389	0.000
$\beta_{(\text{Income})}$	0.014	0.001	24.775	0.000
$\beta_{(\text{Presence of children})}$	0.216	0.011	19.467	0.000
$\beta_{(\text{Distance to public transport})}$	-0.455	0.004	-111.233	0.000

How to interpret the estimate values?

Linear regression – interpretation of parameters

- Linear regression is linear in parameters β
  - Model interpretation: Change of one unit in the independent variable results in β change to the dependent variable, all others being equal
  - Example 1: If a household increases its size by one member then 0.418 daily trips are added
  - Example 2: One additional km far from public transport reduces 0.455 daily trips are subtracted
  - Example 3: If a household has kids then 0.216 daily trips are added.
    - How about households without kids??

Linear regression – interpretation of parameters

- Linear regression is linear in parameters β
  - Model interpretation: Change of one unit in the independent variable results in β change to the dependent variable, all others being equal
  - Example 1: If a household increases its size by one member then 0.418 daily trips are added
  - Example 2: One additional km far from public transport reduces 0.455 daily trips are subtracted
  - Example 3: If a household has children then 0.216 daily trips are added.
    - How about households without kids??
    - Presence of kids is a dummy variable. No kids is a the reference category, no need to estimate a parameter
    - The reference category parameter is fixed typically assumed to be fixed to 0
    - The parameter of the reference category is absorbed by the model constant and the parameters of the categories that we estimate
    - This extends to categorical variables with any number of categories

Linear regression – Estimation output

Using any linear regression software we would receive an output like this:

- But what is the t-ratio? And what is the p-value?

We know what are these…

What is this??

And this??

Parameter	Estimate	s.e.	t-ratio	p-value
β_0	0.883	0.047	18.655	0.000
$\beta_{(\text{Household size})}$	0.418	0.004	97.635	0.000
$\beta_{(\text{Number of vehicles})}$	0.263	0.005	53.389	0.000
$\beta_{(\text{Income})}$	0.014	0.001	24.775	0.000
$\beta_{(\text{Presence of children})}$	0.216	0.011	19.467	0.000
$\beta_{(\text{Distance to public transport})}$	-0.455	0.004	-111.233	0.000

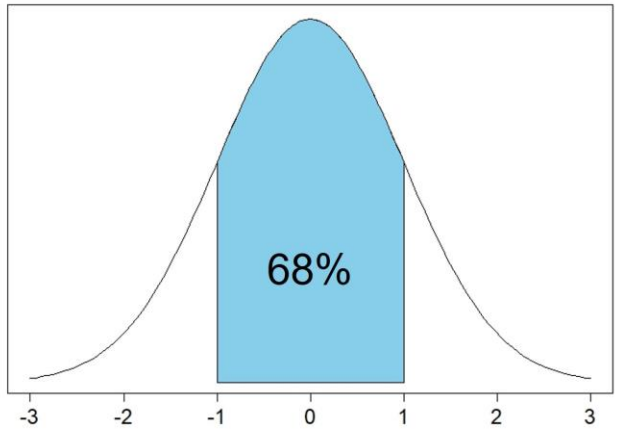
How to interpret the estimate values?

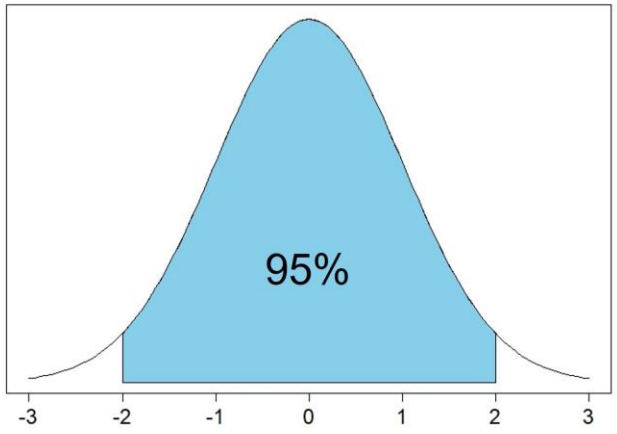
Sampling – Confidence intervals (Reminder)

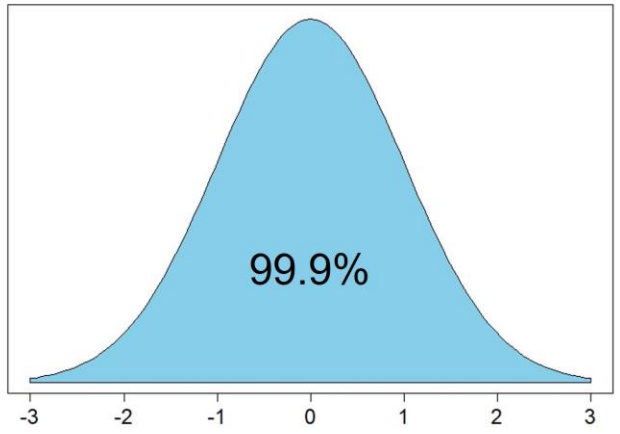
First, let's remember some properties of the standard normal distribution

N ~ (0,1):

- 68% of the observations are between -1 and 1 standard deviations of the mean
- 95% of the observations are between -2 and 2 standard deviations (-1.96 and 1.96 to be precise)
- 99.9% of the observations are between -3 and 3 standard deviations







Sampling – Confidence intervals (Reminder)

- Brief reminder – A standard normal [N ~ (0,1)] variable Z is defined as:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

- A standard normal variable Z is with 0.95 probability between the range [-1.96, 1.96] (from the previous slide); then:

$$0.95 = P\left(-1.96 < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < 1.96\right) = P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

- The confidence interval that captures μ with a probability of 0.95 can be rewritten as:

$$\overline{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

Sampling – Confidence intervals

- For a confidence interval of value (1-α) [where α takes values between 0 and 1] and $Z_{\alpha/2}$ such as the area in each of the two tails of the normal distribution curve (α/2) we get:

$$\overline{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- When population variance is unknown the t-distribution is used instead with degrees of freedom n −1:

$$\overline{X} \pm t_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Sampling – Confidence intervals

	α						
v	**0.1**	**0.05**	**0.025**	**0.01**	**0.005**	**0.0025**	**0.001**
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
35	1.306	1.690	2.030	2.438	2.724	3.340	3.591
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
45	1.301	1.679	2.014	2.412	2.690	3.281	3.520
50	1.299	1.676	2.009	2.403	2.678	3.261	3.496
100	1.290	1.660	1.984	2.364	2.626	3.174	3.390
>100	1.282	1.645	1.960	2.326	2.576	3.091	3.291

Hypothesis testing

Hypothesis testing is used to assess if a difference in a population parameter (e.g. mean) between two or more groups is likely to have occurred by chance or due to some specific factor.

Mechanics of hypothesis testing:

- **Null Hypothesis** (*H*0)**:** There is no significant difference between two groups
- **Alternative Hypothesis** (*H*1)**:** There is significant difference between two groups

Hypothesis testing

Example: We want to investigate whether after the implementation of traffic calming measures, the average traffic speed on a road is different from 60 km/h. Hence:

- H0: $\mu_{speed} = 60km/h$
- H1: $\mu_{speed} \neq 60km/h$

First we transform speed (e.g. we take speed observations from vehicles in the study area) to a Z-variable as:

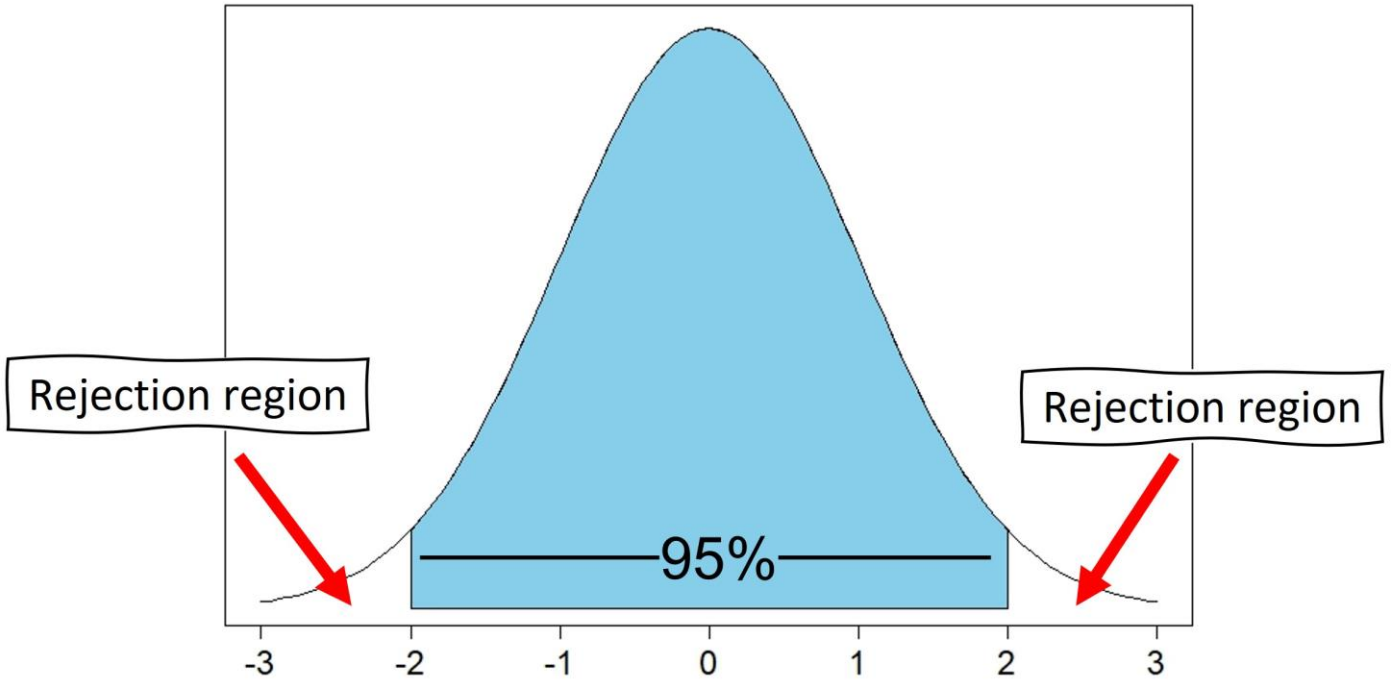
$$Z^* = \frac{\bar{X} - \mu}{\sigma\sqrt{n}}$$

The result of this transformation is a variable of approximately normal distribution with mean = 0 and standard deviation = 1

Hypothesis testing

The null hypothesis is rejected if the sample mean is significantly different from 60 and $\bar{X}$ falls in the rejection region

- The value of the true mean is not within [-1.96, 1.96] with 0.95 probability



Hypothesis testing

Critical Points of Z_c for Different Levels of Significance α

Level of Significance α		
0.10	**0.05**	**0.01**
± 1.645	± 1.960	± 2.576

We then evaluate significance as:

- Critical values of Z, or Z_c, are defined such that

$$P[Z^* \geq Z_c] = P[Z* \leq -Z_c] = \alpha/2$$

- If $|Z^*| \geq |Z_c|$, then the probability of observing this (Z^*) value (or larger), if H0 is true is α. In this case, the null hypothesis (H0) is rejected.
- If $|Z^*| < |Z_c|$, then the probability of observing this value (or smaller) if H0 is true is equal to $1-\alpha$. In this case, the null hypothesis (H0) cannot be rejected.

Hypothesis testing: the p-value

- Probability value or *p*-value
- An alternative metric to report the significance of an outcome
- The smallest level of significance α that leads to rejection of the null hypothesis
- Quantifies the amount of statistical evidence that supports the alternative hypothesis
- Let's say we obtain $Z^* = 3.27$; the *p*-value is calculated as:

$$p\text{-}value(Z^* = 3.27) = p[Z \leq -3.27 \text{ and } Z \geq 3.27] = 2p[Z \geq 3.27] = 2[1-p[Z \leq 3.27]]$$

$$= 2[1-.99946] = .001$$

- For reference, if $Z^* = 1.96$, then *p*-value = 0.05 (typically we want to see *p*-value ≤ 0.05)
- But how do we know that $p[Z \leq 3.27] = .99946$??

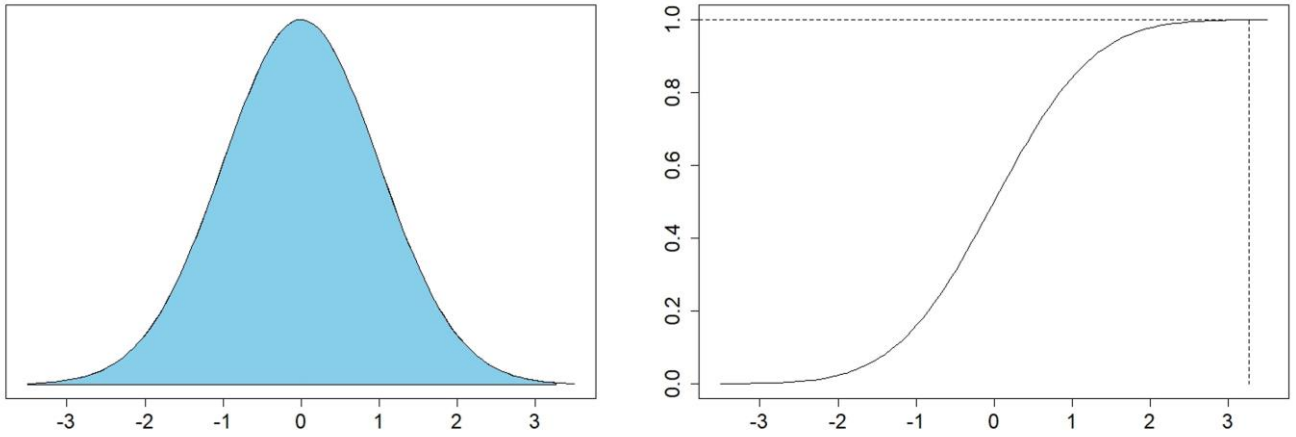
Hypothesis testing: the p-value

- To answer this question we need to use the concept of Cumulative Density Function (CDF)
- The CDF of a probability distribution contains the probabilities that a random variable X is smaller than or equal to a given value x

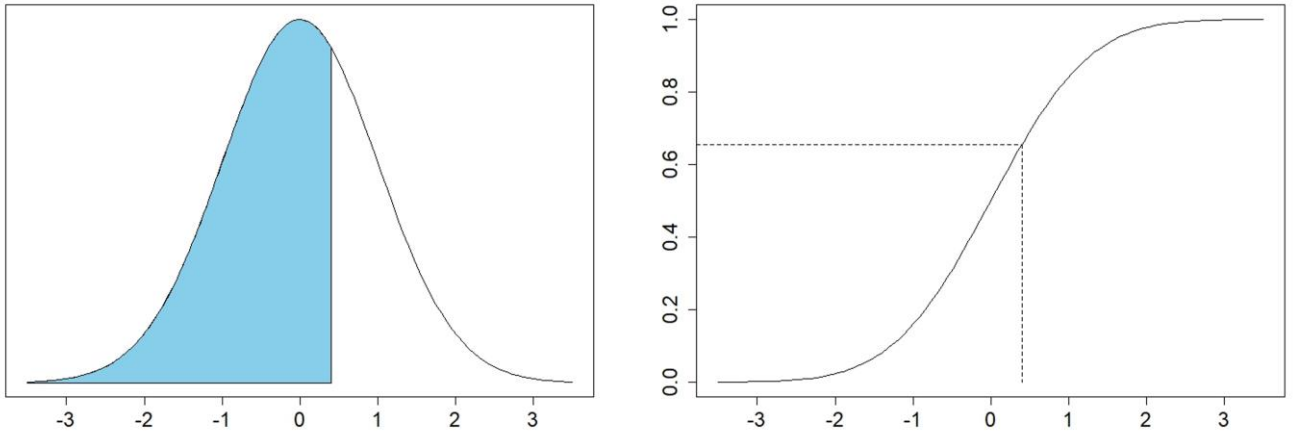
$$F_X(x) = P[X \leq x] = \int_{-\infty}^{x} f_X(x)dx$$

- For the normal distribution, the CDF value for 3.27 is .99946
- CDF functions are readily available in most software packages

Cumulative Density Function (CDF)



$F_X(3.27) = 0.99946$



$F_X(0.40) = 0.65542$

Hypothesis testing and parameter estimates

- In linear regression (and all statistical models) we want to estimate parameters
- $\beta : <\beta_0,\beta_1,\beta_2,...,\beta_N>$ that best approximate our dependent variable
- We need a metric that will allow us to measure the importance of each parameter i.e., the likelihood that an independent variable is significantly related to the dependent.
- It is possible to achieve so with hypothesis testing.

Hypothesis testing and parameter estimates

- Let's focus on the parameter β_1 of a model.
- We can only obtain an estimate of the true β_1; let's call it $\hat{\beta}_1$
- The sampling distribution of the estimate $\hat{\beta}_1$ of β_1 is the distribution of the mean values that would result from repeated samples drawn from the population.
- The sampling distribution is approximately normal (from the Central Limit Theorem) as:

$$\hat{\beta}_1 \approx (\beta_1, \sigma_1^2)$$

where σ_1^2 is the standard error of the parameter

- We can then form a hypothesis test around the true value of β_1

Hypothesis testing and parameter estimates

- The typical hypothesis test that we form is around the 0 value

Why? If the true value of a parameter is not significantly different from 0, then the independent variable associated with this parameter does not have an impact on the dependent variable

- Hence, we have:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

- Following the hypothesis testing approach we discussed earlier, we form the t-statistic of the parameter as:

$$t_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - 0}{\sigma_{\hat{\beta}_1}}$$

- As we already discussed, for large samples we reject the null hypothesis for $|t_{\hat{\beta}_1}| \geq 1.96$ (or $p-\text{Value} < 0.05$)

Linear regression – Estimation output

Now we can understand the whole table

We know what are these…

What is this??

And this??

Parameter	Estimate	s.e.	t-ratio	p-value
β_0	0.883	0.047	18.655	0.000
$\beta_{(\text{Household size})}$	0.418	0.004	97.635	0.000
$\beta_{(\text{Number of vehicles})}$	0.263	0.005	53.389	0.000
$\beta_{(\text{Income})}$	0.014	0.001	24.775	0.000
$\beta_{(\text{Presence of children})}$	0.216	0.011	19.467	0.000
$\beta_{(\text{Distance to public transport})}$	-0.455	0.004	-111.233	0.000

How to interpret the estimate values?

Linear regression Assumptions – Brief summary

Statistical Assumption	Mathematical Expression
1. Functional form	$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$
2. Zero mean of disturbances	$E[\varepsilon_i] = 0$
3. Homoscedasticity of disturbances	$\mathrm{VAR}[\varepsilon_i] = \sigma^2$
4. Nonautocorrelation of disturbances	$COV[\varepsilon_i, \varepsilon_j] = 0$ if $i \neq j$
5. Uncorrelatedness of regressor and disturbances	$COV[X_{1i}, \varepsilon_j] = 0$ for all I and j
6. Normality of disturbances	$\varepsilon_i = N(0, \sigma^2)$

Linear regression – Goodness-of-fit

- Sum of square errors:

$$\text{SSE} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

- Total sum of squares:

$$\text{SST} = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$$

- The coefficient of determination, R-squared (proportion of total variance explained by X)

$$R^2 = 1 - \frac{SSE}{SST}$$

$0 \leq R^2 \leq 1$, if $R^2 = 1$ then all the variance is explained by the independent variables

Linear regression – Useful transformations

Exponential

$Y=\alpha\beta^{X}\varepsilon \Longrightarrow \ln Y=\ln\alpha+X\ln\beta+u$, where $u\sim N(0, \sigma^2)$

Logarithmic (log-log model)

$Y=e^{\alpha}X^{\beta}\varepsilon \Longrightarrow \ln Y=\alpha+\beta\ln X+u$, where $u\sim N(0, \sigma^2)$

Other

$Y = \frac{1}{\alpha+\beta X+u} \Longrightarrow \frac{1}{Y} = \alpha + \beta X + u$, where $u\sim N(0, \sigma^2)$

Linear regression – Comments

- Remember: Significance of parameters and goodness-of-fit are not everything!

- Always check the sign of your parameters:
  - E.g., does it make sense in my study if my model predicts that car availability or household size reduce the number of trips?
  - Make sure that the interpretation of your parameter estimates is consistent with your expectations

Scaling methods

- Models **DO NOT** guarantee that (system closure):

$$\sum_i O_i = \sum_j D_j$$

- We need a closed system to generate an OD matrix
- We assume that generation models are ‘better’ than trip attraction models
  - Generation (production) models: Sophisticated household-based models including explanatory variables
  - Attraction models: Estimated using zonal data
- Fix: total number of trips arising from summing all origins O_i is the ‘correct’ value
  - Each destination D_j are multiplied by an *F* factor as:

$$F_D = \frac{\sum_i O_i}{\sum_j D_j}$$

Scaling methods

- If we have reason to trust more the destination data, then we multiply each O_i by:

$$F_O = \frac{\sum_j D_j}{\sum_i O_i}$$

- If we trust both then:

1. $G = \left(\sum_i O_i + \sum_j D_j\right)/2$
2. $F_O = \frac{G}{\sum_i O_i}$
3. $F_D = \frac{G}{\sum_j D_j}$
4. $O'_i = O_i \, F_O$
5. $D'_j = D_j \, F_D$

Limitations of trip generation models

- Uncertain growth: We have no forecasting information. Define several scenarios an evaluate their plausibility
- Definition of trip generation: Be clear if the total demand changes or we simply observe modal shift or change in route choice in our models
- Unclear behaviour: Non home-based trips are difficult to be modelled well
- Model scope: The standard 4–step model assumes that no one is moving house or work location due to transport issues

Summary

Trip generation modelling (step 1 of the 4–step model)

- Aim – motivation – purpose
- Terminology
- Models
  - Cross classification – category analysis
  - Growth factor models
  - Linear regression