CIVIL-557
Introduction - Fundamentals of statistical modelling

Evangelos Paschalidis

Transport and Mobility Laboratory (TRANSP-OR)
École Polytechnique Fédérale de Lausanne (EPFL)

Contact details

Evangelos Paschalidis, Postdoctoral Researcher
Choice modelling, statistics, driving behaviour modelling

EPFL ENAC IIC TRANSP-OR
Office GC B2 401
evangelos.paschalidis@epfl.ch

I am happy to see students to discuss course problems and issues

This course: Decision-aid methodologies in transportation

Main objective: Demonstrate practical examples and applications of theoretical knowledge in the transport field

- Applied rather than theoretical
- Draws heavily on computer practicals – uses transport datasets (driving behaviour)
- First lecture introduction and theoretical background but then balance between lectures and practicals
- Model estimation scripts implemented in Python
- Script templates will be provided before each lecture

Structure of the course

Teaching:

- Lectures (Theory)
- Practicals (Exercises and coding)

Sometimes, lectures and practicals may be mixed

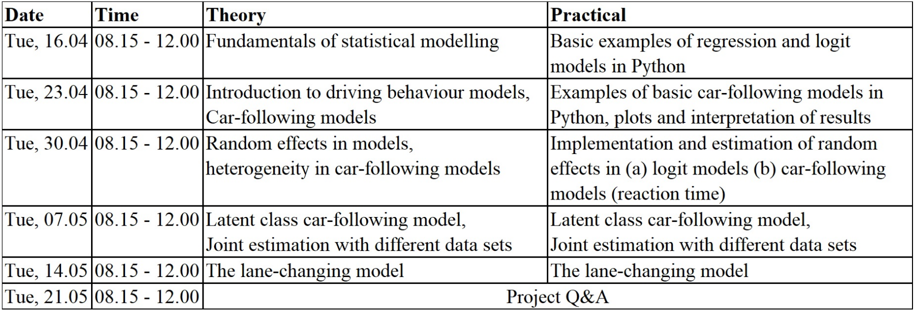
Evaluation:

- Project (50% of the grade - Submission: June 21st)
- Written exam (50% of the grade - May 28th)

Suggested reading:

- Course slides and scripts unless otherwise indicated

Time-table (tentative/will-may change)



Date	Time	Theory	Practical
Tue, 16.04	08.15 - 12.00	Fundamentals of statistical modelling	Basic examples of regression and logit models in Python
Tue, 23.04	08.15 - 12.00	Introduction to driving behaviour models, Car-following models	Examples of basic car-following models in Python, plots and interpretation of results
Tue, 30.04	08.15 - 12.00	Random effects in models, heterogeneity in car-following models	Implementation and estimation of random effects in (a) logit models (b) car-following models (reaction time)
Tue, 07.05	08.15 - 12.00	Latent class car-following model, Joint estimation with different data sets	Latent class car-following model, Joint estimation with different data sets
Tue, 14.05	08.15 - 12.00	The lane-changing model	The lane-changing model
Tue, 21.05	08.15 - 12.00	Project Q&A	

Overview

What is the relevance of statistics and modelling to transportation?

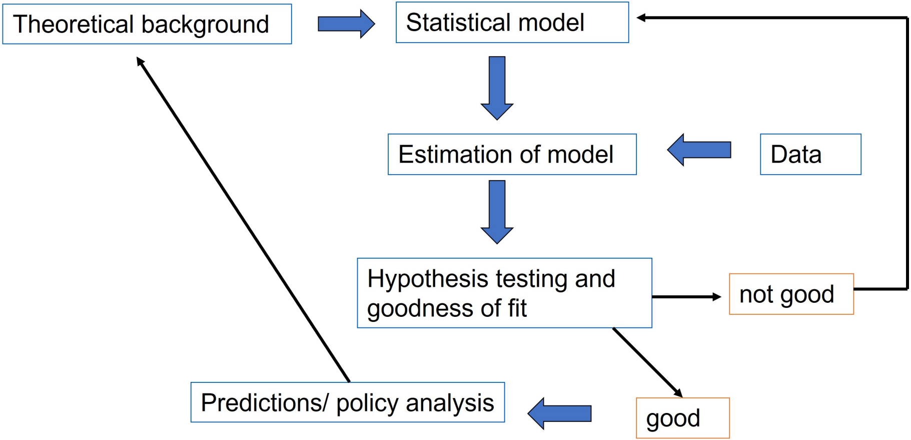
(Let's see some examples...)

Statistical models

Quantitative measurement and analysis of actual phenomena:

1. Describe a phenomenon (Specification of a relationship between a dependent and a series of independent variables)
2. Test a hypothesis (Evaluation of alternative theories with quantitative evidence)
3. Forecasting (Prediction under different scenarios)

Statistical models



Statistical models

Modelling process - Steps

1. Theoretical background - Hypotheses development: The researcher is developing a theoretical framework to test based on existing knowledge.
2. Statistical model (Model specification): The functional form of the model [$y = f(x)$]. Typically describes the relationship between a variable under investigation (dependent variable: y) with a series of other variables (independent or explanatory variables: x).
3. Data / data collection: The data to use for the estimation of the model. These typically come from field observations or questionnaires.
4. Model estimation: The process of calibrating the relationship between the dependent and the independent variables.
5. Hypothesis testing: Evaluate whether the independent variables have a considerable (significant impact) on the dependent variable.
6. Model interpretation: Examine the direction of the relationship between the independent and dependent variables.

Application examples

Example 1: Trip generation

- Understanding trip generation is crucial for accurately predicting traffic patterns and demand for transportation services.
- *A trip generation model* can:
  - Provide insights about the most important factors that affect trip generation
  - Provide predictions for trip generation per household
- Transportation planners can make more informed decisions about infrastructure investments, traffic management, and sustainable transportation solutions, ultimately leading to more efficient and equitable transportation systems.

Application examples

Example 2: Mode choice

- Mode choice models are essential for transportation planning and policy-making.
- *Mode choice models* can:
  - Provide insights about the most influential factors in mode choice
  - Forecast demand for current and future solutions
  - Can inform the design and evaluation of transportation policies (e.g. sustainable mobility)
  - Help to evaluate the effectiveness of existing transportation policies and inform future policies

Application examples

Example 3: Road safety

- *Car crash occurrence and severity models* are essential for improving road safety and reducing the number of traffic accidents and fatalities.
- Understand the factors that contribute to car crashes (e.g. driver behavior, vehicle characteristics, road conditions, and environmental factors).
- Can inform the design and evaluation of road safety interventions, such as improving road infrastructure, increasing public awareness campaigns, and enforcing traffic laws.

Application examples

Example 4: Policy making in railways

- *Railway econometrics* is essential for understanding the economic aspects of railway transportation, such as pricing, demand, and investment.
- Can support efficient and sustainable railway operations and help to allocate resources and prioritize investments.
- Can inform the development of new business models and revenue streams for railways, such as freight transportation and station commercialisation.

Application examples

Example 5: Driving behaviour (Main focus on the course)

- We model longitudinal an latitudinal behaviour of drivers at the individual level
- Typical examples:
    - Car-following (acceleration decisions with respect to the behaviour of a lead vehicle)
    - Lane-changing (decision making process with respect to lane-changing behaviour)
    - Gap acceptance (driver behaviour at intersections, roundabouts etc.)
- Implementation of models in traffic simulation models

Lecture overview

Fundamentals of statistical modelling

- Basic formulations
- Assumptions - estimation
- Evaluation - Hypothesis testing
- Examples

Regression analysis

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

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where

Y_i: Dependent variable

X_i: Independent variable

β_0 & β_1: parameters to be estimated

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

Regression analysis

Advantages: Easy to specify, estimate and interpret.

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

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

Regression assumptions

Assumption I: The dependent variable 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:

$$Trips_i = 0.883 + 0.418 \times \text{Household size}_i - 0.455 \times \text{PT proximity}_i$$

Interpretation:
– One additional family member adds 0.418 trips per day
– One additional km far from public transport reduces daily trips by 0.455

Regression assumptions

- Assumption I is only relevant for linear regression specification strictly.
- We can have linearity in parameters of transformed dependent/independent variables:

$$LnY_i = \beta_0 + \beta_1 LnX_i + \epsilon_i$$

- We can add non-linearities such as:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$$

- We can consider non-additive model specifications:

$$Y_i = \beta_0 \frac{X_{1i}^{\beta_1}}{X_{2i}^{\beta_2}} + \epsilon_i$$

⚠ As long as the ϵ_i is i.i.d. normally distributed, we can still use the principles of normal distribution to estimate models with continuous dependent variable.

⚠ The linear effect of independent variables only holds for linear regression models.

Regression assumptions

Assumption II: The error term ϵ_i has a zero mean:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

$$E(Y_i|X_i) = \beta_0 + \beta_1 X_i$$

$$E(\epsilon_i|X_i) = 0$$

Interpretation: The mean of ϵ does not have any effect on Y

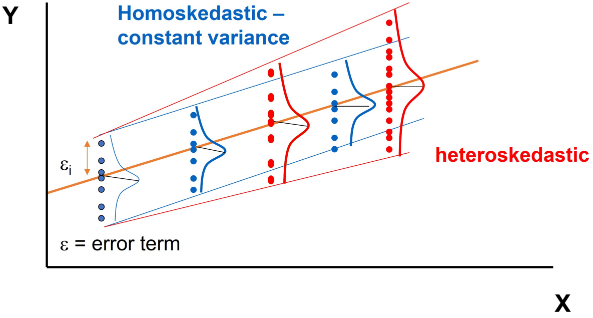
Implication: Biased parameter estimates and standard errors

Regression assumptions

Assumption III: Homoskedasticity:

Variance of the error term $\text{var}(\epsilon_i) = \sigma^2$ is constant across observation and uncorrelated to the independent variables

- Heteroskedasticity can bias the reliability of the significance tests of the parameters



Regression assumptions

Assumption IV: No autocorrelation of the error terms:

The error terms across different observations are not correlated: cov($\epsilon_i \epsilon_j$) $= 0 \; \forall i \neq j$

Implication: There is no correlation between the Ys, cov($Y_i Y_j$) $= 0$

Can affect the reliability of both parameter estimates and standard errors

Assumption V: No correlation between the error term and the independent variables:

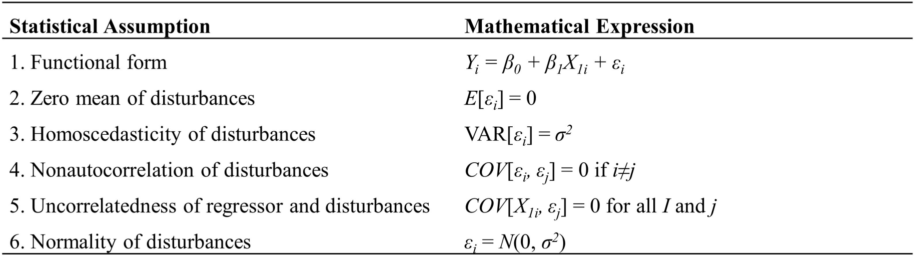
cov($\epsilon_i X_i$) $= 0$ for all X variables

Implication: If cov($\epsilon_i X_i$) $\neq 0$, then the parameter estimates of βs are biased as some influence of ϵ may be attributed to X

Regression assumptions

Assumption VI: The error terms are approximately normally distributed: The distribution of the error terms is $\epsilon \sim N(0, \sigma^2)$ This property allows for the inference of significance of the parameter estimates.

Summary of regression assumptions:



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	$\varepsilon_i = N(0, \sigma^2)$

A short note on dummy 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

A short note on dummy variables

Let's consider a categorical variable 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 in that case would be:

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

In this case we keep D_4 as a reference category and we are not estimating a parameter for it.

What if the dependent variable is not continuous?

The dependent variable can have several different formats e.g:

- Continuous: Speed of a vehicle
- Ordered: A survey response to a 5-level Likert scale question
- Discrete: Mode choice for commuting
- Count: Number of crashes per year

– Linear regression is not a suitable model for all types of dependent variable.

– The choice of a suitable model depends on the nature of the dependent variable.

Fundamentals of logit models

Logit models are the most typical models to use when the dependent variable is discrete (e.g. mode choice)

- Dependent variable: A discrete outcome that indicates the presence of a condition
- Outcome: Population proportion or probability *(P)* for the occurrence of the selected outcome
- The concept of utility theory is used in many transportation applications:
  - Alternatives (conditions) have a utility based on attributes (e.g. travel time or travel cost of a transport mode)
  - Utility affects behaviour (choice or presence of condition)

Fundamentals of logit models

Framework of a logit model:

- A decision maker n faces J alternatives, with $j = 1, 2, \ldots, J$
- Every alternative is described by a set of K attributes $x_{n,j} = \langle x_{n,j,1}, \ldots, x_{n,j,K} \rangle$
- A vector of preference weights (parameters) $\beta_{n,k}$ are associated with $x_{n,j,k}$
- Utility is then specified as
  $U_{n,j} = V_{n,j} + \epsilon_{n,j} =$
  $U_{n,j} = \beta_{n,1} x_{n,j,1} + \ldots + \beta_{n,K} x_{n,j,K} + \epsilon_{n,j}$

Fundamentals of logit models

- The error terms ($\epsilon_{n,j}$):
  - Are distributed independently across alternatives and individuals [same as linear regression]
  - Are distributed identically across alternatives and individuals [same as linear regression]
  - Follow a type-I extreme value distribution (Gumbel) with location parameter η and scale parameter μ
- The variance of the distribution is $\frac{\pi^2}{6\mu^2}$
- In logit models, we normalise η to 0 and μ to 1

Fundamentals of logit models

- Choice probability:

$$P_{n,i} = P(\mu V_{n,i} + \mu\epsilon_{n,i} \geq \mu V_{n,j} + \mu\epsilon_{n,j}, \forall j \neq i)$$

- Under the properties of Gumbel distribution:

$$P_{n,i} = \frac{e^{\mu V_{n,i}}}{\sum_{j=1}^{J} e^{\mu V_{n,j}}}$$

- The scale μ cannot be estimated but is normalised to 1:

$$P_{n,i} = \frac{e^{V_{n,i}}}{\sum_{j=1}^{J} e^{V_{n,j}}}$$

Fundamentals of logit models

Choices are not affected by the:

- Addition of the same constant to all utilities
- Multiplication of all utilities by the same (positive) constant

– Only differences in utilities matter

– This is the reason scale is normalised to 1

– One of the alternative specific constant variables must also be normalised to 0

– Parameter estimates between different logit models cannot be directly estimated because the actual scale is unknown (and the results that we estimate are in fact $\mu\beta$)

Estimation procedures

- We have now specified a model [either continuous (e.g. linear regression) or discrete (e.g. logit)]...
- Our model specification includes M unknown parameters $\langle\beta_0, \beta_1, ..., \beta_{m-1}\rangle$to be estimated
- We need parameter values that best represent our observed outcomes (dependent variable)
- Following the maximum likelihood estimation (MLE) procedure we need to:
  1. Specify the probability (likelihood) of a dependent variable observation as a function of independent variables and parameters to be estimated
  2. Find the values of parameters β that maximise the total likelihood

Maximum likelihood estimation (MLE)

- In practice, we maximise the log-likelihood (LL) rather than the likelihood (for computational convenience).
- The LL function as a function of β is specified as:
  $$LL(\beta) = \sum_{n=1}^{N} ln(P_{n,j_n}(\beta))$$
- For the estimation of parameters, we usually use gradient based algorithms such as Berndt–Hall–Hall–Hausman (BHHH) or Broyden-Fletcher-Goldfarb-Shanno (BFGS).

MLE example for linear regression

- Let's consider the following linear regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

  with $Y \sim N(\mu, \sigma^2)$.

- For continuous variables we can calculate the probability density function (PDF) of the observations.
- The PDF for normally distributed variables is

$$f(Y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-0.5(\frac{Y-\mu}{\sigma})^2}$$

- Following the MLE approach, we need to find the value μ that will maximise the PDF $f(Y)$ of our dependent variable y observations.

MLE example for linear regression

For N observations, the linear regression model is:

$$\begin{gathered}
Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \epsilon_1 \\
Y_2 = \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \epsilon_2 \\
\vdots \\
Y_N = \beta_0 + \beta_1 X_{1N} + \beta_2 X_{2N} + \epsilon_N
\end{gathered}$$

MLE example for linear regression

After replacing the regression model in the density function we get:

$$f(Y_1) = \frac{1}{\sigma\sqrt{2\pi}} e^{-0.5(\frac{Y_1-(\beta_0+\beta_1 X_{11}+\beta_2 X_{21})}{\sigma})^2}$$
$$f(Y_2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-0.5(\frac{Y_2-(\beta_0+\beta_1 X_{12}+\beta_2 X_{22})}{\sigma})^2}$$
$$\vdots$$
$$f(Y_N) = \frac{1}{\sigma\sqrt{2\pi}} e^{-0.5(\frac{Y_N-(\beta_0+\beta_1 X_{1N}+\beta_2 X_{2N})}{\sigma})^2}$$

MLE example for linear regression

– The joint probability (density) of all observations (assuming observation independence) is

$$L(\beta) = \prod_{i=1}^{N} (f_i(\beta))$$

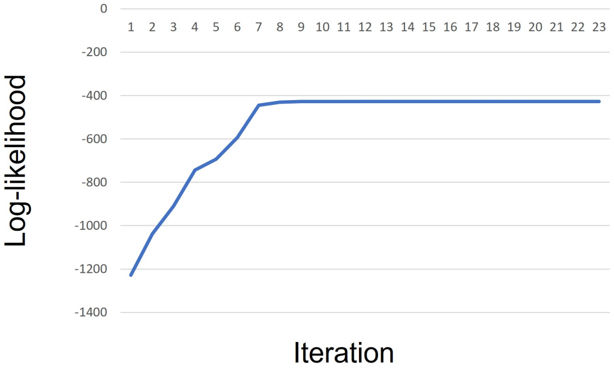
– As we said, instead of the probability (density) function we prefer to maximise the logarithm *ln*, hence

$$LL(\beta) = \sum_{i=1}^{N} ln(f_i(\beta))$$

** For continuous variables, the probability density (rather than actual probability) is used, as probability cannot be defined for continuous outcomes.*

Log-likelihood progress

- Sharper steps in the initial iterations of the estimation
- We expect LL to be a negative number; probabilities take values (0,1) hence their *ln* is negative
- Probability density function can take values > 1, it may happen that LL is positive for such models



What about non-linear and other model specifications?

- For continuous variables, as long as the error term ϵ is additive and i.i.d. normally distributed, the normal density function can still be used.
- For other (non-normal) distributions of error terms we must use the density function of the respective distribution of the error term.
- For logit model, the LL estimation procedure is identical with regression models but we must replace the density function with the logit probability: $P_{n,i} = \frac{e^{V_{n,i}}}{\sum_{j=1}^{J} e^{V_{n,j}}}$

But how to evaluate our results?

We have now implemented MLE and obtained results (values for the β coefficients)

How do we evaluate these results?

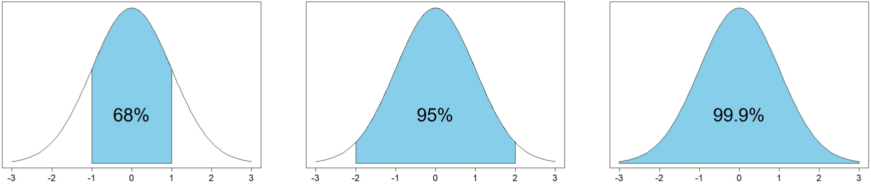
First let's introduce a few statistical tools:

- Confidence intervals (CIs)
- Hypothesis testing

Confidence intervals

First, let's remember some properties of the standard normal distribution $N \sim (0,1)$:

- 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



Confidence intervals

Central limit theorem: When a sufficiently large random sample is drawn from a population with mean μ and standard deviation σ, the sample mean $\overline{X}$ is approximately normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$.

Now, keeping in mind that a standard normal variable Z is with 0.95 probability between the range [-1.96, 1.96] (from the previous slide) we derive:

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

Confidence intervals

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

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

Other probability levels, such as 0.90 can be used, too.

$$\overline{X} \pm 1.645 \frac{\sigma}{\sqrt{n}}$$

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 ($\alpha/2$) we get:

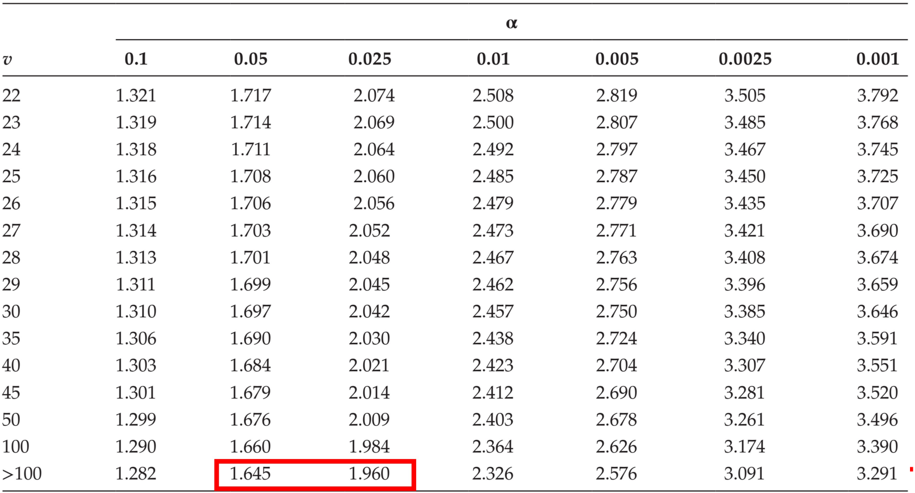
$$\overline{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

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

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

Confidence intervals

The *t*-distribution table



	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:

- H_0: $\mu_{speed} = 60km/h$
- H_1: $\mu_{speed} \neq 60km/h$

First we need to transform speed to a Z-variable as:

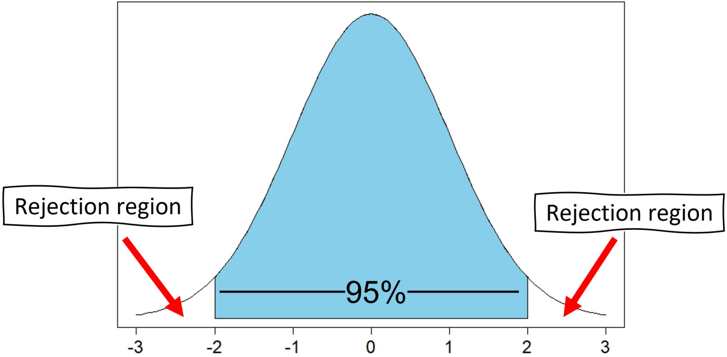
$$Z^* = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

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

EPFL

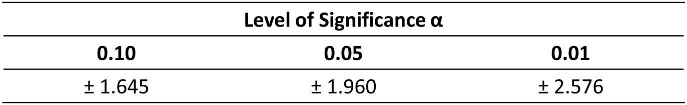
Hypothesis testing

⚠ **Reminder:** The null hypothesis is rejected if the sample mean is significantly different from 60 and $\overline{X}$ falls in the rejection region



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 value (or larger), if H_0 is true is α. In this case, the null hypothesis (H_0) is rejected.
- If $|Z^*| < |Z_c|$, then the probability of observing this value (or smaller) if H_0 is true is equal to $1 - \alpha$. In this case, the null hypothesis (H_0) cannot be rejected.

Hypothesis testing and $p-$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:

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

$$= 2p[Z \geq 3.27] = 2[1 - p[Z \leq 3.27]]$$

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

Cumulative Density Function

??? But how do we know that $p[Z \leq 3.27] = .99946$???

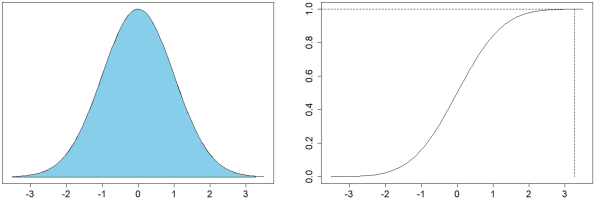
- To answer this question we need to use the concept of Cumulative Density Function (CDF)
- The CDF of a probability distribution contains the probabilities that a random variable X is less than or equal to a given value x

$$F_X(x) = P[X \leq x] = \int_{-\infty}^{x} f_X(x)\,dx, \text{ for all } x \in \mathbb{R}$$

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

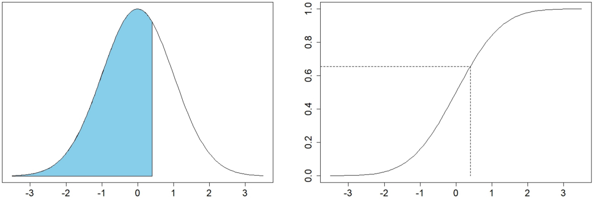
Cumulative Density Function

Illustration of PDF and CDF for $x = 3.27$



$F_X(3.27) = 0.99946$

Illustration of PDF and CDF for $x = 0.40$



$F_X(0.40) = 0.65542$

And all these are relevant because..??

- In statistical models we want to estimate parameter values $\beta : <\beta_0, \beta_1, \beta_2, ..., \beta_N>$ 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.

Inference in statistical models

- Let's focus on the parameter β_1^* of a model.
- The sampling distribution of the estimate B_1 of β_1^* is the distribution of values that would result from repeated samples drawn from the population.
- The sampling distribution is approximately normal as:

$$B_1 \approx N(\beta_1^*, \sigma_{\beta_1}^2)$$

- σ_{β_1} is also known as the standard error of the parameter.
- We can then form a hypothesis test around the true value of ${\beta_1}^*$.

Inference in statistical models

- 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:

$$H_0 : B_1 = 0$$
$$H_1 : B_1 \neq 0$$

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

$$t_{B_1} = \frac{B_1 - 0}{\sigma_{\beta_1}}$$

- As we already discussed, for large samples we reject the null hypothesis for $|t_{\beta_1}| \geq 1.96$ (or $p-$Value < 0.05)

Some additional notes on hypothesis testing

- H_0: null hypothesis
- H_1: alternative hypothesis

		Decision	
		Do not reject H_0	Reject H_0
Truth	H_0 True		*Type I error probability α*
	H_0 False	*Type II error probability β*	

$\beta = 1 - \alpha$

Basic goodness-of-fit statistics

- ρ^2 measure (bounded between 0 and 1)

$$\rho^2 = 1 - \frac{LL(\overline{\theta})}{LL(0)}$$

- if $\rho^2 = 1$ then perfect fit
  - There is something suspicious that needs further investigation
- Not to be mixed with the linear regression coefficient of determination R^2 which is also between 0 and 1

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}} = 1 - \frac{\sum(y_i - fi)^2}{\sum(y_i - \overline{y})^2}$$

Basic goodness-of-fit statistics

- ρ^2 is dependent on:
  - the sample data and size
  - the number of estimated parameters
  - the number of discrete outcomes (in logit models)
- It can be used to compare models only when the requirements above are met
- ρ^2 has a risk of over-fitting model to the data
- the adjusted ρ^2 $(\overline{\rho^2})$ can be used instead as:

$$\rho^2 = 1 - \frac{LL(\overline{\theta}) - K}{LL(0)}$$

- Useful to compare models with different number of parameters
- Penalisation of increase of parameters

EPFL

The AIC and BIC tests

- AIC and BIC are both used for model selection among a finite set of models.
- They help in choosing the model that best balances the goodness of fit and complexity.

Akaike Information Criterion (AIC)

- Formula: $\mathrm{AIC} = 2k - 2\ln(L)$
- Here, k is the number of parameters in the model, and L is the maximum value of the likelihood function for the model.
- Lower AIC values indicate a better model.

Bayesian Information Criterion (BIC)

- Formula: $\text{BIC} = \ln(n)k - 2\ln(L)$
- In this formula, n is the number of observations or sample size, k is the number of parameters, and L is the likelihood.
- Like AIC, a lower BIC value suggests a better model, but it penalizes complexity more strongly.

Summary of modelling procedure

1. Decide the model type e.g., regression model, logit model etc.
2. Model specification
   - Define the relationship between parameters and independent variables
   - Define the probability (or probability density function) functional form
3. Evaluate parameter estimates
   - Parameter signs are consistent with expectations
   - Significance of parameters (normally t-ratio > 1.96 or $p-$Value < 0.05)

Lecture overview

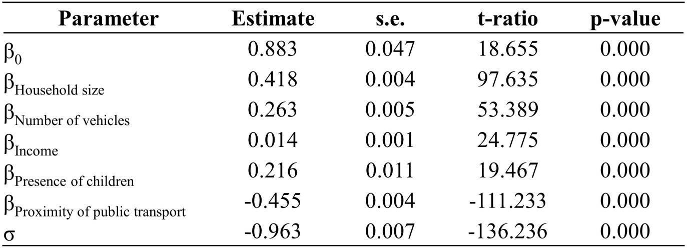
Fundamentals of statistical modelling

- Basic formulations
- Assumptions - estimation
- Evaluation - Hypothesis testing

Questions??

Example model outputs

Linear regression model of trip generation - Dependent variable: Number of trips per household

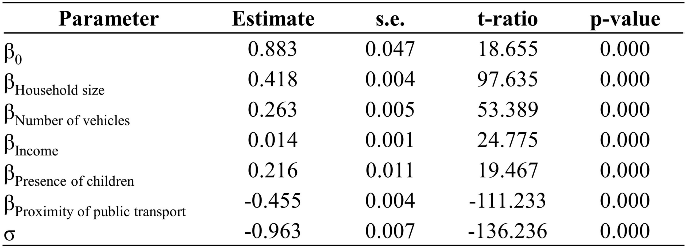


Parameter	Estimate	s.e.	t-ratio	p-value
β_0	0.883	0.047	18.655	0.000
$\beta_{\text{Household size}}$	0.418	0.004	97.635	0.000
$\beta_{\text{Number of vehicles}}$	0.263	0.005	53.389	0.000
β_{Income}	0.014	0.001	24.775	0.000
$\beta_{\text{Presence of children}}$	0.216	0.011	19.467	0.000
$\beta_{\text{Proximity of public transport}}$	-0.455	0.004	-111.233	0.000
σ	-0.963	0.007	-136.236	0.000

Example model outputs

Linear regression model of trip generation - Dependent variable: Number of trips per household

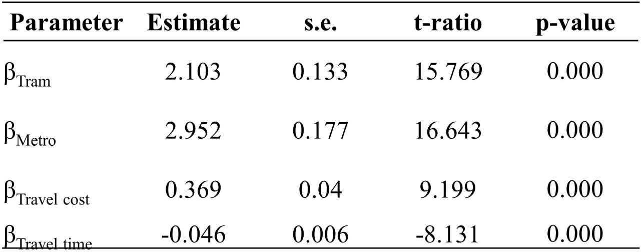
- What can we say about the significance of the results?
- Are parameter estimates consistent with our expectations?



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
β_{Income}	0.014	0.001	24.775	0.000
$\beta_{\text{Presence of children}}$	0.216	0.011	19.467	0.000
$\beta_{\text{Proximity of public transport}}$	-0.455	0.004	-111.233	0.000
σ	-0.963	0.007	-136.236	0.000

Example model outputs

Multinomial logit model of mode choice - Dependent variable: Choice between bus, metro and tram



Parameter	Estimate	s.e.	t-ratio	p-value
β_{Tram}	2.103	0.133	15.769	0.000
β_{Metro}	2.952	0.177	16.643	0.000
$\beta_{\text{Travel cost}}$	0.369	0.04	9.199	0.000
$\beta_{\text{Travel time}}$	-0.046	0.006	-8.131	0.000

Reminder: The parameter estimates in logit model affect the utility for choosing one alternative over the others

Model comparison: Example 1

Assume two model specifications:

Model 1:
$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon_1$

Model 2:
$y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3 + \gamma_4 X_4 + \epsilon_2$

Can we compare the fit of the two models using the likelihood-ratio test? If not then which test would you choose?

Model comparison: Example 2

Assume two model specifications:

Model 1:
$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon_1$

Model 2:
$y = \gamma_0 X_1^{\gamma_1} X_2^{\gamma_2} X_3^{\gamma_3} + \epsilon_2$

Can we compare the fit of the two models using the likelihood-ratio test? If not then which test would you choose?

Practical examples - Lab

- Linear regression model: Trip generation
- Linear regression model: Car-following
- Multinomial logit model: Mode choice
- Binary logit model: Crowdshipping route acceptance