

Random utility and the logit model

Stephane Hess

stephane.hess@gmail.com

Random utility and the logit model

Outline

- ① Utility theory
- ② MNL model: background and model structure
- ③ Specifying utility functions

$$\lambda^x e^{-\lambda} \sum_{x=0}^{\infty} P(x) = 1$$

Utility theory

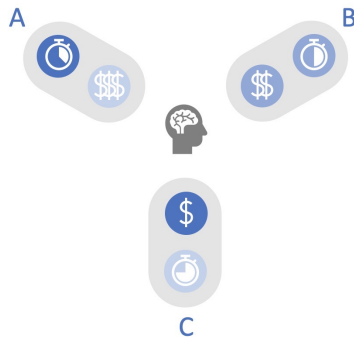
$$W = \pi \int f^2(x) dx$$

$$\frac{1}{\sqrt{1-k^2 \sin^2(x)}} = 1 + \sum_{n=1}^{\infty} \sum_{x=0}^{\infty} \frac{(2n)!}{(2n)!}$$

Utility theory

General concept of utility maximisation

- ❑ Alternatives characterised by utility
 - based on attributes which influence behaviour
- ❑ Assume rational behaviour
 - choose alternative with highest utility
- ❑ Trade-off behaviour:
 - good performance on one attribute compensates for poor performance on another
 - e.g. higher cost compensated by faster journey



Utility theory

Utility specification

Decision-maker

- Person n , with $n = 1, \dots, N$
- Faces T_n choice situations, with $t = 1, \dots, T_n$
- Characteristics z_n (observed)
- Vector of preferences/tastes β_n (estimated)

Choice-set and context

- J mutually exclusive alternatives, with $j = 1, \dots, J$
- Choice context described by w_{nt}
- Alt. j described by set of K attributes
- In situation t , $x_{jnt} = \langle x_{jnt,1}, \dots, x_{jnt,K} \rangle$

$$U_{jnt} = f(\beta_n, x_{jnt}, w_{nt}, z_n)$$

Utility theory

Utility and choice

- For now, drop indices n and t
- Choice index given by Y
- Alternative with highest utility is chosen

$$Y = i \iff U_i > U_j \quad \forall j \neq i$$

- Observation:
only differences in utility matter

$$\begin{aligned} U_j^* &= U_j + \Delta \quad \forall j \\ Y^* &= Y \quad \forall \Delta \end{aligned}$$

Utility theory

A simple example

- Choice between two train services
- Alternatives can only be distinguished via their attributes
- Both attributes are continuous
- Use a linear in attributes specification
- If T_1 increases by one minute, U_1 changes by β_T
- Expect β_T and β_C to be negative

Unlabelled choice situation

	Train 1	Train 2
Travel time (T)	45 min	30 min
Travel cost (C)	£7	£12

Utility specification

$$U_1 = \beta_T T_1 + \beta_C C_1$$

$$U_2 = \beta_T T_2 + \beta_C C_2$$

Utility theory

Choice outcome

- Choice depends on differences in utilities
- If sensitivity to time increases, differences in time matter more (same for cost)
- If all sensitivities increase by same factor, order of preferences does not change

Utility specification

$$U_1 = \beta_T T_1 + \beta_C C_1$$

$$U_2 = \beta_T T_2 + \beta_C C_2$$

Choice outcome

$$Y = 1 \iff \beta_T (T_1 - T_2) > \beta_C (C_2 - C_1)$$

$$\beta_T, \beta_C < 0 \Rightarrow \frac{\beta_T}{\beta_C} (T_2 - T_1) > C_1 - C_2$$

Utility theory

Example for our choice task: $T_1 > T_2$ and $C_1 < C_2$

- ❑ Option 1 will be chosen if $\frac{\beta_T}{\beta_C} < \frac{C_2 - C_1}{T_1 - T_2}$
 - not willing to pay extra cost to save time
- ❑ Option 2 will be chosen if $\frac{\beta_T}{\beta_C} > \frac{C_2 - C_1}{T_1 - T_2}$
- ❑ No information from observations with *dominant* alternative
- ❑ To find β values, need many observations with changing attribute levels

Choice scenario

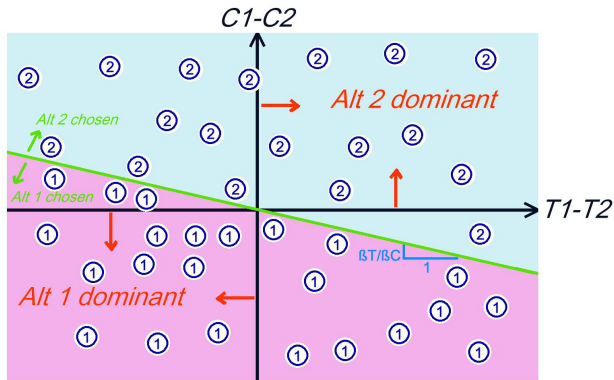
	Train 1	Train 2
Travel time (T)	45 min	30 min
Travel cost (C)	£7	£12

Choice outcome

$$Y = 1 \iff \frac{\beta_T}{\beta_C} < \frac{C_2 - C_1}{T_1 - T_2}$$

Utility theory

We can solve this graphically



Utility theory

Shortcomings of deterministic utility theory

- ❑ Often make observations of
 - inconsistent behaviour
 - non-transitive preferences
- ❑ Cause of inconsistencies cannot be specified in deterministic framework
 - lack of analyst's knowledge of individual's decision processes
 - unobserved attributes
 - unobserved heterogeneity
 - incorrectly measured attributes
 - poor information on availabilities
 - non-linearities in preferences
- ❑ To accommodate this, we move to a probabilistic model

Utility theory

Random utility theory

- Utility U_{jn} is a random variable
 - *deterministic* part V_{jn}
 - *random* part ε_{jn}
- Deterministic part specified to capture role of observed explanators
- ε_{jn} measures deviation from modelled utility for alternative j and respondent n

Additive utility structure

$$U_{jn} = V_{jn} + \varepsilon_{jn}$$

Deterministic part of utility

$$V_{jn} = f(\beta_n, x_{jn})$$

Utility theory

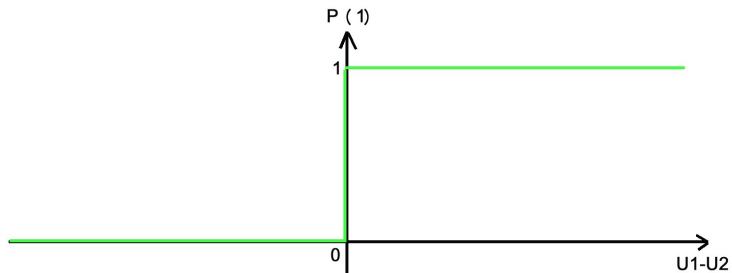
Implications for probabilities

- Deterministic utility theory
 - Alternative with highest utility is chosen
- Random utility theory
 - Probability of choosing alternative increases with deterministic utility
- Probability of person n choosing alternative i given by:

$$P_{in} = P(V_{in} + \varepsilon_{in} > V_{jn} + \varepsilon_{jn} \quad \forall j \neq i)$$

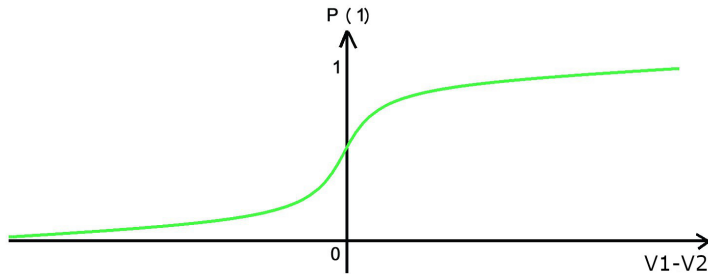
Utility theory

Binary deterministic choice



Utility theory

Binary probabilistic choice



Utility theory

Only differences in utility matter

- Probability of person n choosing alternative i given by:

$$P_{in} = P(V_{in} + \varepsilon_{in} > V_{jn} + \varepsilon_{jn} \quad \forall j \neq i)$$

- Adding same value to $V_{jn} \quad \forall j$ will not change probabilities
- Observation: only differences in utilities matter
- Implication: parameters only identified if they capture differences across alternatives
 - require normalisation for e.g. alternative specific constants

Utility theory

Overall scale of utility is irrelevant

- Multiplication of $U_j \forall j$ by $\lambda > 0$ yields equivalent model

$$P_{in} = P(V_{in} + \varepsilon_{in} > V_{jn} + \varepsilon_{jn} \quad \forall j \neq i)$$

$$P_{in}^* = P(\lambda \cdot V_{in} + \lambda \cdot \varepsilon_{in} > \lambda \cdot V_{jn} + \lambda \cdot \varepsilon_{jn} \quad \forall j \neq i)$$

- Increases in V can be counteracted by increases in the variance of ε , and vice versa
- Observation: overall scale of utility is irrelevant
- Implication: we need to normalise the overall scale of utility
- Standard solution: normalise the variance of the error terms

Utility theory

Level of randomness

Probability: $P_{in} = P(V_{in} + \varepsilon_{in} > V_{jn} + \varepsilon_{jn} \quad \forall j \neq i)$

- ❑ multi-dimensional integral over $f(\varepsilon_n)$
- ❑ closed form only for certain choices of $f(\varepsilon_n)$
- ❑ degree of randomness depends on relative size of V and ε

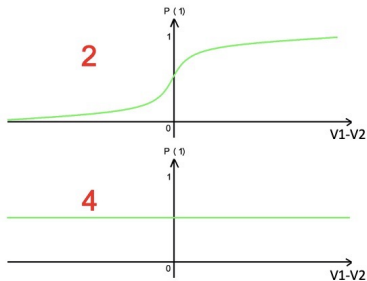
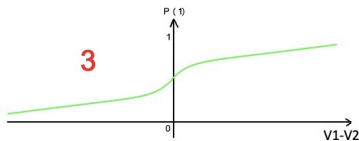
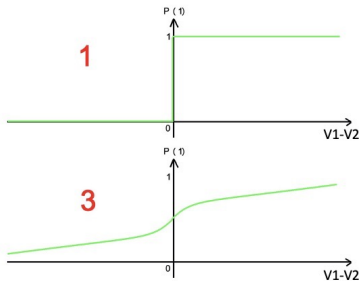
Concept of scale of model

- ❑ inversely proportional to variance of error
- ❑ higher scale means more deterministic choice process
- ❑ higher relative weight for modelled utility
 - with variance of ε normalised, higher scale would mean larger coefficients

Utility theory

Different degrees of error

- *Randomness* increases with uncertainty
- Move from step function to $P_j = \frac{1}{J}$



MNL model: background and model structure

MNL model: background and model structure

Introduction

History and role

- ❑ Used since 1960s
- ❑ Most basic of random utility models
- ❑ But should still be your starting point

Conditional logit or *multinomial* logit?

- ❑ “Conditional logit” originally used for model that uses characteristics of alternatives
- ❑ “Multinomial logit” used for model using decision maker characteristics
- ❑ In practice, we do both at the same time
- ❑ I use the term Multinomial Logit (MNL)

MNL model: background and model structure

Core assumptions

Structure of a model depends on assumptions about error term

$$U_{jn} = V_{jn} + \varepsilon_{jn}$$

- MNL uses a *iid* type I extreme value (EV) distribution

Homoskedasticity

- ε distrib. **identically** across alternatives and respondents
⇒ extent of noise is the same

Independence

- ε distrib. **independently** across alternatives and respondents
⇒ no correlation across alternatives

Type I EV distribution

- Parameters: η and μ
- Mean: $\eta + \frac{\gamma}{\mu}$
 - γ : Euler const. (~ 0.577)
- Variance: $\frac{\pi^2}{6\mu^2}$

MNL model: background and model structure

Type I extreme value distribution

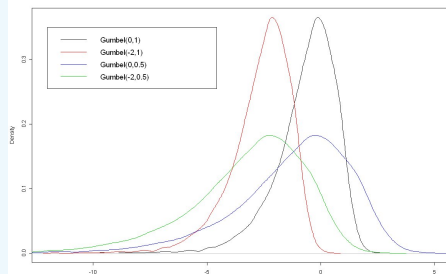
Density function

$$f(\varepsilon) = \mu e^{-\mu(\varepsilon-\eta)} e^{-e^{-\mu(\varepsilon-\eta)}}$$

Cumulative distribution function

$$F(\varepsilon) = e^{-e^{-\mu(\varepsilon-\eta)}}$$

Shape of distribution



MNL model: background and model structure

Decisions on η and μ

- Normalise mean of error terms by setting $\eta_j = 0, \forall j$
 - mean of $\varepsilon_{jn} = \frac{\gamma}{\mu} \forall j, n$, where γ is Euler constant (~ 0.577)
 - not zero, but only differences in utility matter
- For iid errors, use generic μ
 - $\text{var}(\varepsilon_{jn}) = \frac{\pi^2}{6\mu^2}, \forall j$
 - note that this means we have not yet *normalised* the variance

MNL model: background and model structure

Choice probabilities

- With $\varepsilon \sim EV1$, obtain closed form solution for probabilities

$$P_{in} = P(V_{in} + \varepsilon_{in} > V_{jn} + \varepsilon_{jn} \quad \forall j \neq i)$$

$$P_{in} = \frac{e^{\mu V_{in}}}{\sum_{j=1}^J e^{\mu V_{jn}}}$$

- Denominator is across all alternatives, so $0 \leq P_{jn} \leq 1, \forall j$ and $\sum_{j=1}^J P_{jn} = 1$
- μ is the scale parameter, with $\text{var}(\varepsilon_{jn}) = \frac{\pi^2}{6\mu^2}, \forall j$
- Shows us the limiting cases of MNL
 - $\lim_{\mu \rightarrow 0} P_{jn} = \frac{1}{J}, \forall j$
 - $\lim_{\mu \rightarrow \infty} P_{in} = 1$ if $V_{in} = \max(V_{1n}, \dots, V_{Jn})$

MNL model: background and model structure

Implications for scale

Variance and scale

- Increasing μ means lower variance for error term
- $\lim_{\mu \rightarrow 0} P_{jn} = \frac{1}{J}, \forall j$
- $\lim_{\mu \rightarrow \infty} P_{in} = 1$ if $V_{in} = \max(V_{1n}, \dots, V_{Jn})$

$$\text{var}(\varepsilon_{jn}) = \frac{\pi^2}{6\mu^2}, \forall j$$

Overspecification and need for normalisation

- Increasing all parameters in V has the same impact as increasing μ
 - See [MNL_probabilities.xlsx](#) (sheet 1)
- Can ensure that model is identified by applying a normalisation
 - usual choice is to set $\mu = 1$
 - easier/more general than fixing one β

MNL model: background and model structure

Normalising the variance of ε

- Set $\mu = 1$

$$\text{var}(\varepsilon_{n,j}) = \frac{\pi^2}{6}, \forall j$$

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

- See [MNL_probabilities.xlsx](#) (sheets 2 and 3)

$$\lambda^x e^{-\lambda} \sum_{x=0}^{\infty} P(x) = 1$$

Specifying utility functions

Specifying utility functions

Utilities and alternatives

- ❑ V_i depends ONLY on the attributes of alternative i
- ❑ Attributes of alternative j or even existence of another alternative should not enter V_i
- ❑ Ensures consistency with utility maximisation
- ❑ Violation of this rule may lead to preference reversals
- ❑ Probability of choosing i of course still depends on the existence and attributes of j
- ❑ Choice set effects (e.g. impact of presence of product j) should be captured in the model structure (i.e. error structure), not the utilities

Specifying utility functions

Recap on random utility maximisation (RUM)

- RUM theory recognises that analyst cannot fully explain utility
- Typically use an additive structure

$$\begin{aligned}U_{jn} &= V_{jn} + \varepsilon_{jn} \\V_{jn} &= f(\beta, x_{jn}, z_n, w_n) \\ \varepsilon_n &\sim f(\varepsilon_n), \text{ with } \varepsilon_n = \langle \varepsilon_{1n}, \dots, \varepsilon_{Jn} \rangle\end{aligned}$$

- For each model, make assumptions about the mean and covariance of error terms

Specifying utility functions

Assumption about equal means for error term

- Typical assumption:

$$\begin{aligned} E(\varepsilon_{in}) &= E(\varepsilon_{jn}), \forall i, j \\ \Rightarrow E(\varepsilon_{in}) - E(\varepsilon_{jn}) &= 0, \forall i, j \end{aligned}$$

- No systematic differences between alternatives captured in the error term



Specifying utility functions



The “all else equal” assumption

- Choice between two car journeys
- Trade-off between time and cost
- Basic utility specification:

$$V_{car_1} = \beta_t \cdot T_1 + \beta_c \cdot C_1$$
$$V_{car_2} = \beta_t \cdot T_2 + \beta_c \cdot C_2$$

- What if $T_1 = T_2$ and $C_1 = C_2$?
 - would expect equal choice shares
 - if no other factors affect choice!



		
time (T, minutes)	25	20
cost (C, £)	3	4

		
time (T, minutes)	25	25
cost (C, £)	4	4

Specifying utility functions





“Point of indifference” example

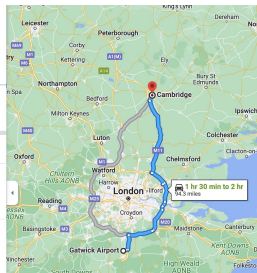
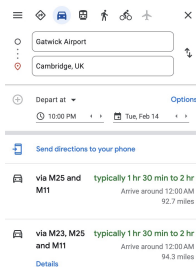
- ❑ Trade-off between time and cost
- ❑ Point of indifference
 - what if £1 is as bad as 5 minutes?
 - would have $V_{car_1} = V_{car_2}$
 - would obtain equal choice shares
 - again, only if no other factors affect the difference

		
time (T, minutes)	25	20
cost (C, £)	3	4

Specifying utility functions

Not all differences are captured by attributes in V

		
time (T, minutes)	25	20
cost (C, £)	3	4
		
time (T, minutes)	60	60
cost (C, £)	5	5



Specifying utility functions

Possible sources of differences

- ❑ Unobserved effects with a non-zero mean
 - Labels or types of alternatives
 - modal preferences (e.g. car vs train) or brand preferences (e.g. Apple vs Samsung)
 - status quo or opt-out alternatives
 - Properties of alternatives
 - cheap vs expensive
 - Positioning/presentation effects
 - order effects (e.g. left vs right, top vs bottom)
- ❑ The *true* mean impacts of unobserved factors may vary across alternatives
- ❑ It is then dangerous to specify a model with $E(\varepsilon_{in}) = E(\varepsilon_{jn}), \forall i, j$
- ❑ Could lead to bias in other parameters

Specifying utility functions

Inclusion of alternative specific constants (ASCs)

- Utility function with ASC:

$$U_{jn} = \delta_j + f(\beta, x_{jn}) + \varepsilon_{jn}$$

(subject to a normalisation)

- ASCs capture mean of unobserved part of utility
- Can now assume that $E(\varepsilon_{in}) = E(\varepsilon_{jn})$, $\forall i, j$

Specifying utility functions



Benefits of ASCs

- ❑ Means of unobserved effects may differ across alternatives
- ❑ Model without ASCs:
 - Potential bias in β
 - Reduced model fit
 - Wider confidence intervals
- ❑ ASCs also help recovery of market shares

Specifying utility functions

Normalisation of ASCs

- Only differences in utility matter
- With J alternatives, can estimate $J - 1$ constants
- In our example: $\delta_{car} = -\delta_{train}$
- Choice of base is arbitrary if ASCs are not random
- See [MNL_probabilities.xlsx](#) (sheet 4)

		
time (minutes)	25	20
cost (£)	3	4

$$V_{car} = \delta_{car} + \beta_t \cdot T_{car} + \beta_c \cdot C_{car}$$

$$V_{train} = \beta_t \cdot T_{train} + \beta_c \cdot C_{train}$$

$$V_{car} = \beta_t \cdot T_{car} + \beta_c \cdot C_{car}$$

$$V_{train} = \delta_{train} + \beta_t \cdot T_{train} + \beta_c \cdot C_{train}$$

Specifying utility functions

Not all attributes are continuous

- ❑ Wifi attribute can take three levels:
 - $W=1$: no wifi
 - $W=2$: Available for £1
 - $W=3$: Available for free
- ❑ Is a linear specification still appropriate?

$$\begin{aligned}V_B &= \delta_B + \beta_T T_B + \beta_C C_B + \beta_{wifi} W_B \\V_T &= \beta_T T_T + \beta_C C_T + \beta_{wifi} W_T\end{aligned}$$

	Bus	Train
Travel time (T)	45 min	30 min
Travel cost (C)	£7	£12
Wifi	Available for £1	Available for free

Specifying utility functions

Problems with a linear specification

	Bus	Train
Travel time (T)	45 min	30 min
Travel cost (C)	£7	£12
Wifi	Available for £1	Available for free

- What does $\beta_{wifi} W_B$ (respectively $\beta_{wifi} W_T$) imply?
- Assumes linear and monotonic effect
 - Going from $W_B = 1$ to $W_B = 2$ is the same as from $W_B = 2$ to $W_B = 3$
 - Going from $W_B = 1$ to $W_B = 3$ is twice as good as going from $W_B = 1$ to $W_B = 2$

Specifying utility functions

A categorical approach again needs a normalisation

- How about?

$$\begin{aligned} V_B &= \dots + \beta_{wifi_1} (W_B == 1) + \beta_{wifi_2} (W_B == 2) + \beta_{wifi_3} (W_B == 3) \\ V_T &= \dots + \beta_{wifi_1} (W_T == 1) + \beta_{wifi_2} (W_T == 2) + \beta_{wifi_3} (W_T == 3) \end{aligned}$$

- This is again an over-specification
- Only differences in utility matter
- Infinite number of combinations of β_{wifi_1} , β_{wifi_2} and β_{wifi_3} will give same differences

Specifying utility functions

Dummy vs effects coding

Dummy coding

- ❑ Set the parameter for one category to zero, e.g. $\beta_{wifi_3} = 0$
- ❑ Other parameters will measure differences in utility across categories

Effects coding

- ❑ Set the parameter for one category to be the negative sum of the others, e.g. $\beta_{wifi_3} = -\beta_{wifi_1} - \beta_{wifi_2}$
- ❑ Third parameter calculated, not estimated

Do they give the same results?

- ❑ For many years, view persisted that dummy coding was affected by confounding
- ❑ Daly et al. (2016) show this is not true

Key reference: Daly, A.J., Dekker, T. & Hess, S. (2016), *Dummy coding vs effects coding for categorical variables: clarifications and extensions*, *Journal of Choice Modelling*, 21, pp. 36-41.

Specifying utility functions

Moving away from linearity

- ❑ Is 4% risk of infection only 8 times as bad as 0.5% risk?
- ❑ Is 5 years of protection 2.5 times as good as 2 years?
- ❑ Empirical evidence that marginal utilities are function of the level of an attribute
 - i.e. $\frac{\partial V_i}{\partial x_{i,k}}$ is not independent of the value of $x_{i,k}$
- ❑ Also happens with cost, even if decreasing marginal sensitivity is counterintuitive

Please consider the following vaccination options and make your choice as if they happened in the current environment. Please remember there is no right or wrong answer.

	Vaccine A	Vaccine B	No vaccine
Risk of infection (out of 100,000 people coming in contact with infected person):	500 (0.5%)	4,000 (4%)	7,500 (7.5%)
Risk of serious illness (out of 100,000 people who become infected):	4,000 (4%)	2,000 (2%)	20,000 (20%)
Estimated protection duration:	two years	five years	
Risk of mild side effects (out of 100,000 vaccinated people):	5,000 (5%)	100 (0.1%)	
Risk of severe side effects (out of 100,000 vaccinated people):	1 (0.001%)	1 (0.001%)	
Population coverage:	40%		
Exemption from international travel restrictions:	exempt		
Waiting time (free vaccination):	6 months	6 months	
Fee (no waiting time):	E400	E400	

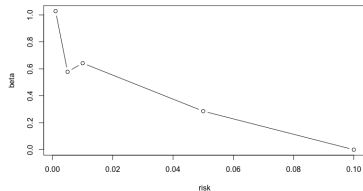
	Vaccine A free	Vaccine A paid	Vaccine B free	Vaccine B paid	No vaccine
Yourself:	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Specifying utility functions

Brute force approach

- ❑ Treat continuous attributes as categorical
- ❑ Quickly run into problems
 - We may have too many levels to do this
 - Estimates may not be nicely monotonic
 - But often clear evidence of non-linearity

risk of mild side effects	estimate	rob t
0.1	0	NA
0.05	0.28587	2.9914
0.01	0.64208	6.0968
0.005	0.57798	6.0318
0.001	1.02794	8.4572

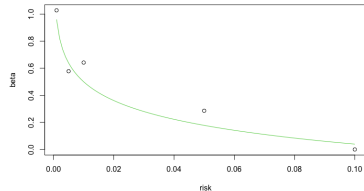
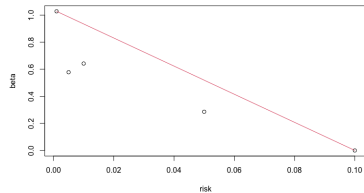


Specifying utility functions

Linear vs non-linear

- ❑ Not accounting for non-linearity can cause bias, also in unaffected parameters
 - e.g. datasets where linear cost assumption leads to positive estimate for time coefficient
- ❑ Common to impose a specific level of non-linearity, e.g. replace βx with $\beta \ln(x)$
- ❑ Better to estimate level of non-linearity

Box-Cox transform: $\beta \frac{x^\lambda - 1}{\lambda}$, with $\lambda > 0$
polynomial specification: $\beta x + \beta_2 x^2 + \dots$



Summary

Summary

Key points from this class

- ❑ Utility theory is key paradigm in choice modelling
- ❑ Utility of an alternative depends only on attributes of that alternative
- ❑ RUM acknowledges limited ability of analyst to understand decision process
- ❑ Multinomial Logit (MNL) is the base model, and should always be our starting point
- ❑ Specification of utility is key step, requires care and understanding of identification

Summary

Suggested reading

- Train, K.E. (2009), Discrete Choice Methods with Simulation, Cambridge University Press, free online access <https://eml.berkeley.edu/books/choice2.html>
 - Chapter 3



Questions?



www.ApolloChoiceModelling.com

The most flexible choice modelling software (up to a probability)