ENAC
Laboratoire TRANSP-OR

Prof. Michel Bierlaire

Mathematical Modeling of Behavior

January 29^{th} 2024

Name	
Signature	Section

Question	Points	
1	20	
2	20	
3	20	
4	20	

Total		Grade	

This exam is written and lasts 3 hours, from 9:15 to 12:15.

The only material that you are allowed to use is the handwritten summary, maximum length of 4 pages (2 double-sided A4 sheets or 4 single-side A4 sheets).

The summaries will be collected at the end along with the exam.

The exam comprises four questions. Please answer each question in the space provided after it. The TAs can provide you with additional sheets if needed.

Make sure that your name and the date are mentioned on every page of the exam and on the summary sheets.

You shall answer in English.

All answers have to be carefully justified.

No calculator is allowed.
Make sure to simplify your calculations as much as possible.

Question 1

(20 points)
In a simple model developed for the London passenger mode choice (LPMC) dataset, the utilities for the walking (WA), cycling (CY), public transport (PT) and driving (DR) alternatives are specified as follows:

$$V_{in} = \text{ASC}_i + \beta_{\text{time}} \, \text{time}_{in} + \beta_{\text{cost}} \, \text{cost}_{in} + \varepsilon_{in},$$

where time_{in} is the travel time (in hours) and cost_{in} is the travel cost (in GBP) associated with alternative $i \in \{\text{WA}, \text{CY}, \text{PT}, \text{DR}\}$ and individual n, and ε_{in} are error terms that are independently and identically extreme value-distributed:

$$\varepsilon_{in} \overset{\text{iid}}{\sim} \text{EV}\left(0, \mu\right).$$

Note that the walking and cycling alternatives have no associated cost (that is, $\text{cost}_{\text{WA},n} = \text{cost}_{\text{CY},n} = 0, \forall n$). Moreover, μ is normalized to one and ASC_{WA} is normalized to zero. The estimates of the model parameters are given in Table 1.

Name	Value	Std. err.	t-test
μ	1	—	—
ASC_{WA}	0	—	—
ASC_{CY}	−4.20	0.50	
ASC_{PT}	−0.60	0.40	
ASC_{DR}	−1.50	0.30	
β_{time}	−7.10	1.00	
β_{cost}	−0.25	0.02	

Table 1: Parameter estimates

Name	Value
μ	
ASC_{WA}	
ASC_{CY}	
ASC_{PT}	
ASC_{DR}	
β_{time}	
β_{cost}	

Table 2: Estimates, $\text{ASC}_{\text{CY}} = 0$

Name	Value
μ	
ASC_{WA}	
ASC_{CY}	
ASC_{PT}	
ASC_{DR}	
β_{time}	
β_{cost}	

Table 3: Estimates, money-metric

1. **[3 points]** Fill out the values of the t-tests in Table 1. Are all parameter estimates significant at the 95% level? Based on these results, what modeling decision would you take? Why?

2. **[2 points]** Suppose that ASC_{CY} is normalized to zero instead of ASC_{WA}. What values do the parameter estimates take? Fill out Table 2. What happens if one tries to estimate all ASCs?

3. **[2 points]** The estimation results in Table 1 were obtained using the linear-in-parameters normalization. What values would the parameter estimates take if the money-metric normalization was used instead? Fill out Table 3. Compute the value of time.

Consider now the following modifications of the model:

(a) In all alternatives ($i \in \{\text{WA}, \text{CY}, \text{PT}, \text{DR}\}$), replace parameter β_{time} by an alternative-specific parameter $\beta_{\text{time},i}$;

(b) In all alternatives, replace the variable time_{in} by a Box-Cox transformation $B(\text{time}_{in}; \lambda)$ defined as:

$$B(\text{time}_{in}; \lambda) = \begin{cases} \dfrac{\text{time}_{in}^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \log(\text{time}_{in}) & \text{if } \lambda = 0; \end{cases}$$

(c) In the public transport and driving alternatives, *add* a piecewise-linear interaction $W(\beta_{\text{cost}}; \text{inc}_n)$ defined as:

$$\begin{aligned} W(\beta_{\text{cost}}; \text{inc}_n) = \; & [\beta_{\text{cost,0–4k}} \min(\text{inc}_n, 4000) + \\ & \beta_{\text{cost,4k+}} \max(\text{inc}_n - 4000, 0)] \, \text{cost}_{in}, \end{aligned}$$

where inc_n is the monthly income of individual n (in GBP);

(d) In the public transport and driving alternatives, replace parameter β_{cost} by a random parameter $\tilde{\beta}_{\text{cost}} = \bar{\beta}_{\text{cost}} + \sigma\xi_n$, where

$$\xi_n \overset{\text{iid}}{\sim} \mathcal{N}(0, 1).$$

4. **[4 points]** Clearly state the underlying assumption of each modification.

5. **[4 points]** Can each modification be individually tested against the simple model by means of a likelihood ratio test? If yes, write down explicitly the linear restrictions defining the null hypotheses of each test; otherwise, choose and explain a statistical test that could be used instead.

6. **[1 point]** What are the units of $\beta_{\text{cost,0–4k}}$ and $\beta_{\text{cost,4k+}}$?

7. **[1 point]** Mention one drawback of the Box-Cox transformation.

8. **[1 point]** The random parameter $\tilde{\beta}_{\text{cost}}$ follows a normal distribution, which is usually considered inappropriate. Why?

Suppose that some observations from the LPMC dataset are associated with individual choice sets that include fewer alternatives. Namely, N_4 observations come from individuals for whom all four alternatives are available, N_3 observations come from respondents that have access to only three, and so on for N_2 and N_1.

9. **[1 point]** Write the *null* log likelihood as a function of N_1, N_2, N_3 and N_4.

10. **[1 point]** Explain why excluding all observations that have a single alternative available does not affect the maximum likelihood estimates of the model parameters.

CORRECTION

Question 1

(20 points)
In a simple model developed for the London passenger mode choice (LPMC) dataset, the utilities for the walking (WA), cycling (CY), public transport (PT) and driving (DR) alternatives are specified as follows:

$$V_{in} = \text{ASC}_i + \beta_{\text{time}} \, \text{time}_{in} + \beta_{\text{cost}} \, \text{cost}_{in} + \varepsilon_{in},$$

where time_{in} is the travel time (in hours) and cost_{in} is the travel cost (in GBP) associated with alternative $i \in \{\text{WA}, \text{CY}, \text{PT}, \text{DR}\}$ and individual n, and ε_{in} are error terms that are independently and identically extreme value-distributed:

$$\varepsilon_{in} \overset{\text{iid}}{\sim} \text{EV}\left(0, \mu\right).$$

Note that the walking and cycling alternatives have no associated cost (that is, $\text{cost}_{\text{WA},n} = \text{cost}_{\text{CY},n} = 0, \forall n$). Moreover, μ is normalized to one and ASC_{WA} is normalized to zero. The estimates of the model parameters are given in Table 1.

Name	Value	Std. err.	t-test
μ	1	—	—
ASC_{WA}	0	—	—
ASC_{CY}	−4.20	0.50	**−8.4**
ASC_{PT}	−0.60	0.40	**−1.5**
ASC_{DR}	−1.50	0.30	**−5.0**
β_{time}	−7.10	1.00	**−7.1**
β_{cost}	−0.25	0.02	**−12.5**

Table 4: Parameter estimates

Name	Value
μ	**1**
ASC_{WA}	**4.20**
ASC_{CY}	**0**
ASC_{PT}	**3.60**
ASC_{DR}	**2.70**
β_{time}	**−7.10**
β_{cost}	**−0.25**

Table 5: Estimates, $\text{ASC}_{\text{CY}} = 0$

Name	Value
μ	**0.25**
ASC_{WA}	**0**
ASC_{CY}	**−16.80**
ASC_{PT}	**−2.40**
ASC_{DR}	**−6.00**
β_{time}	**−28.40**
β_{cost}	**−1**

Table 6: Estimates, money-metric

1. **[3 points]** Fill out the values of the t-tests in Table 1. Are all parameter estimates significant at the 95% level? Based on these results, what modeling decision would you take? Why?

   - **[1.5]** See *Table 4*
   - **[0.5]** ASC_{PT} *is not significant at the 95% level.*
   - **[1.0]** *Keep the parameter anyway! ASCs should always be kept. The use of ASCs relaxes the assumption that the location parameters are the same across alternatives.*

2. **[2 points]** Suppose that ASC_{CY} is normalized to zero instead of ASC_{WA}. What values do the parameter estimates take? Fill out Table 2. What happens if one tries to estimate all ASCs?

   - **[1.5]** See *Table 5*
   - **[0.5]** *The model cannot be identified. Only the difference in utility matters.*

3. **[2 points]** The estimation results in Table 1 were obtained using the linear-in-parameters normalization. What values would the parameter estimates take if the money-metric normalization was used instead? Fill out Table 3. Compute the value of time.

   - **[1.75]** See *Table 6*
   - **[0.25]** $\text{VoT} = \beta_{\text{time}}/\beta_{\text{cost}} = 28.40\,[\text{GBP/h}]$.

Consider now the following modifications of the model:

(a) In all alternatives ($i \in \{\text{WA}, \text{CY}, \text{PT}, \text{DR}\}$), replace parameter β_{time} by an alternative-specific parameter $\beta_{\text{time},i}$;

(b) In all alternatives, replace the variable time_{in} by a Box-Cox transformation $B(\text{time}_{in}; \lambda)$ defined as:

$$B(\text{time}_{in}; \lambda) = \begin{cases} \dfrac{\text{time}_{in}^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \log(\text{time}_{in}) & \text{if } \lambda = 0; \end{cases}$$

(c) In the public transport and driving alternatives, *add* a piecewise-linear interaction $W(\beta_{\text{cost}}; \text{inc}_n)$ defined as:

$$\begin{aligned} W(\beta_{\text{cost}}; \text{inc}_n) = \; & [\beta_{\text{cost,0–4k}} \min(\text{inc}_n, 4000) + \\ & \beta_{\text{cost,4k+}} \max(\text{inc}_n - 4000, 0)]\, \text{cost}_{in}, \end{aligned}$$

where inc_n is the monthly income of individual n (in GBP);

(d) In the public transport and driving alternatives, replace parameter β_{cost} by a random parameter $\tilde{\beta}_{\text{cost}} = \bar{\beta}_{\text{cost}} + \sigma\xi_n$, where

$$\xi_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1).$$

4. **[4 points]** Clearly state the underlying assumption of each modification.

   (a) **[1.0]** *Travel time is perceived differently in each alternative.*

   (b) **[1.0]** *Marginal effect of travel time varies with travel time.*

   (c) **[1.0]** *Sensitivity to cost varies linearly with income. Change of rate at 4k.*

   (d) **[1.0]** *Sensitivity to cost is normally distributed in the sample.*

5. **[4 points]** Can each modification be individually tested against the simple model by means of a likelihood ratio test? If yes, write down explicitly the linear restrictions defining the null hypotheses of each test; otherwise, choose and explain a statistical test that could be used instead.

   (a) **[1.0]** *Yes.* $\beta_{\text{time,WA}} = \beta_{\text{time,CY}} = \beta_{\text{time,PT}} = \beta_{\text{time,DR}} \, (= \beta_{\text{time}})$.

   (b) **[1.0]** *Yes.* $\lambda = 1$.

   (c) **[1.0]** *Yes.* $\beta_{\text{cost,0–4k}} = \beta_{\text{cost,4k+}} = 0$.

   (d) **[1.0]** *Yes.* $\sigma = 0$.

6. **[1 point]** What are the units of $\beta_{\text{cost,0–4k}}$ and $\beta_{\text{cost,4k+}}$?

   - **[1.0]** $\left[\text{GBP}^{-2}\right]$.

7. **[1 point]** What is the main drawback of the Box-Cox transformation, when it comes to model estimation?

   - **[1.0]** *Estimation is more complex because utilities are not linear in parameters; OR Behavioral interpretation is not straightforward; OR Danger of overfitting.*

8. **[1 point]** The random parameter $\tilde{\beta}_{\text{cost}}$ follows a normal distribution, which is usually considered inappropriate. Why?

   - **[1.0]** *Support of normal distribution is infinite, which means that some individuals get a positive β_{cost}. Also correct: normal distribution is symmetric, which is not associated with any behavioral aspect.*

Suppose that some observations from the LPMC dataset are associated with individual choice sets that include fewer alternatives. Namely, N_4 observations come from individuals for whom all four alternatives are available, N_3 observations come from respondents that have access to only three, and so on for N_2 and N_1.

9. **[1 point]** Write the *null* log likelihood as a function of N_1, N_2, N_3 and N_4.

    - **[1.0]** $\mathcal{L} = N_2 \times \log(\frac{1}{2}) + N_3 \times \log(\frac{1}{3}) + N_4 \times \log(\frac{1}{4})$, *or, equivalently,* $\mathcal{L} = -N_2 \times \log(2) - N_3 \times \log(3) - N_4 \times \log(4)$.

10. **[1 point]** Explain why excluding all observations that have a single alternative available does not affect the maximum likelihood estimates of the model parameters.

    - **[1.0]** *Choice probability of those individuals is independent from the parameter values, always equal to one. There is nothing to maximize!*

Question 2

(20 points)
This question concerns nested logit and is divided into two parts. The parts are independent and can be solved separately from each other.

Part 1 [8 points].
In this question, we will consider a stylized problem of duplicates. Assume that the population consists of 1000 travelers. Each traveler must choose between two transportation alternatives: a car or a blue bus. The sole factor influencing their decision is travel time, which is assumed to be identical for both the car and the bus.

1. **[1.5 points]** As a choice modeling expert, you are asked by a policymaker to evaluate the population's benefit from introducing a new red bus, which has the same travel time as the existing blue bus. To conduct this analysis, you use the logit model, the simplest model applicable in this scenario, with the normalization parameter set to 1. Your task involves calculating the expected gain for the population from the introduction of the red bus, factoring in this new option within the logit model framework. Describe how to do it and provide the formula for the expected gain. [Hint: consider consumer surplus]

2. **[3 points]** The government, eager to expand public transportation, proposes adding several new buses, each with different colors, to the existing fleet of three buses. All buses, new and existing, have the same travel time. Each new bus incurs a cost of 100 units. What is the optimal number of buses for societal benefit? Determine the optimal number, using an objective function that calculates the difference between total surplus (benefit to society) and total cost.

   To ensure a more accurate analysis and to safeguard future reputation, you have persuaded the policymaker to adopt the nested logit model, which is more appropriate for this scenario.

3. **[2.5 points]** In this setting, travelers have the option to choose from three buses or one car, with all travel options having the same travel time. The buses are categorized in one nest, and the car is in a separate nest. We set the scale parameter within the bus nest μ_m at 1 and assume the scale parameter for differences between the two nests μ is 1.5. Consider if such an assumption is appropriate by focusing on two aspects: a) the ratio between the scale parameters (μ_m and μ), and b) the impact of these parameters on the probabilities of choosing either a bus or a car.

4. **[1 point]** Consider a scenario where the correlation between the error terms for different buses is 0.96, corresponding to a scale parameter μ of 0.2. Without actually computing the optimal number of buses using a nested logit model, would you anticipate the number of buses to be greater or lesser than what

would be derived using a logit model? Provide an explanation for your reasoning.

Part 2 [12 points].
Recall that a choice model verifies the Independence from Irrelevant Alternatives (IIA) property if the ratio of the choice probabilities of any pair of alternatives (i, j) in the choice set does not depend on any other alternative than i and j.
Consider a situation in which decision-maker has a choice among three alternatives $C_1 = \{i, j, k\}$. We assume that a choice model that is used satisfies IIA. The model predicts the following probabilities

$$P(i|C_1) = \frac{1}{2}, \quad P(j|C_1) = \frac{1}{3}, \quad P(k|C_1) = \frac{1}{6}.$$

Suppose that alternative k has been removed from the choice set, so the new choice set is $C_2 = \{i, j\}$.

1. **[2.5 points]** Using IIA property derive the choice probabilities i, j in new choice set C_2.

2. **[1.5 points]** Calculate the change in probabilities:
   $\Delta P(i) = P(i|C_2) - P(i|C_1), \Delta P(j) = P(j|C_2) - P(j|C_1)$ and compare the ratio of these changes to the original probability ratio, specifically comparing $\frac{P(i|C_1)}{P(j|C_1)}$ and $\frac{\Delta P(i)}{\Delta P(j)}$. Comment on your findings.

   Imagine a scenario where a decision-maker is faced with four alternatives: $C = \{i, j, k, \ell\}$. Assume that the choice model is a nested logit model, with the nesting structure as depicted in Figure 1.

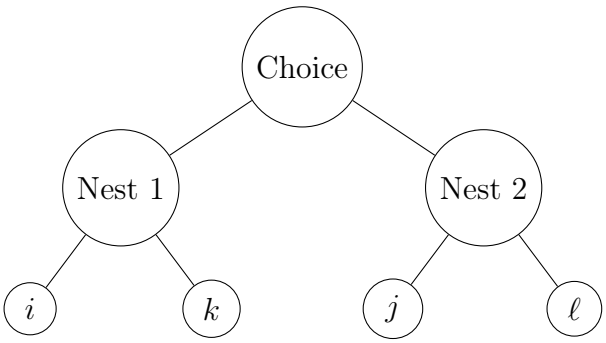


Figure 1: Mode Choice Tree Diagram

3. **[1.5 points]** Are there any pairs of alternatives that satisfy the IIA property? Please list all such pairs.

   In a transportation mode choice survey, you opt to add a supplementary question: "If one of the transport options were unavailable, which mode would you

choose?" The following table displays the aggregate responses to this question. A column reports the choice probabilities of a situation where an alternative is unavailable.

Probabilities					
(Probabilities in parentheses are changes compared to the original)					
Alternative	**Original**	**Auto**	**Walking**	**Bus**	**Rail**
Auto	0.4	-	0.45 (+12.5%)	0.52 (+30%)	0.48 (+20%)
Walking	0.1	0.2 (+100%)	-	0.13 (+30%)	0.12 (+20%)
Bus	0.3	0.48 (+60%)	0.33 (+10%)	-	0.4 (+33%)
Rail	0.2	0.32 (+60%)	0.22 (+10%)	0.35 (+70%)	-

4. **[3 points]** Suppose that you want to use a nested logit model. Does the provided information suggest the structure of the nests? Draw the nest structure and explain your choice. [Hint: IIA property provided above can be used here]

5. **[1 points]** Suppose that you want to test your model against logit. Which test would you use?

6. **[2.5 points]** Your colleague suggests using mixed logit instead of nested logit. Is the mixed logit model capable of capturing correlations among error terms? Please explain your reasoning. If it is capable, outline the model and suggest a method for comparing it with a standard logit model. If not, clarify why it's unsuitable.

CORRECTION

Question 2

(20 points)

Part 1 [8 points].

1. **[1.5 points]** We need to compare the total surplus before and after realizing new bus. In the case of MNL this quantity equals $1000(\ln(3e^{\beta T}) - \ln(2e^{\beta T})) = 1000(\ln 3 - \ln 2)$.

2. **[3 points]** Denoting k is the number of new buses, the objective function of the government equals

$$\max_k \{1000 \ln(3e^{\beta T} + ke^{\beta T}) - 100k\}.$$

Using the property of logarithm we can rewrite the problem as

$$\max_k \{1000 \ln(3 + k) - 100k\}.$$

The objective function is differentiable, therefore, the solution is either on the boundary $k = 0$ or satisfies the foc. The derivative of the function at $k = 0$ equals to $\frac{1000}{3} - 100$ and is positive, therefore, the solution satisfies the first-order condition. The foc results into the equation

$$1000 \frac{1}{k^* + 3} = 100$$

The equation has a unique solution $k^* = 7$. Therefore, it is optimal to build new 7 buses.

3. **[2.5 points]** a) The condition for the the ratio $\frac{\mu}{\mu_m}$ to be less than 1 is violated.

b) The probability of choosing a car equals

$$P(\text{car}) = \frac{e^{\mu\beta T}}{e^{\mu\beta T} + e^{\mu(\beta T + \ln 3)}} = \frac{1}{1 + 3^{\mu}}.$$

If $\mu = 1.5 > 1$, then the probability of choosing a car is less than $\frac{1}{4}$. Therefore, the probability of not choosing a car is greater than $\frac{3}{4}$ and the probability of choosing a bus is greater than $\frac{1}{4}$. Thus, the probability of choosing a bus is greater than a car. This pattern of behavior is not realistic. Moreover, in this case, a nested logit model can not be derived from the maximization of the random utility framework.

4. **[1 point]** Intuitively, that the optimal number of buses is less in the case of the nested logit because travelers should choose a bus less often than the car because of the substitutability of buses.

Part 2 [12 points].

1. **[2.5 points]** From IIA property equality $\frac{P(i|C_1)}{P(j|C_1)} = \frac{P(i|C_2)}{P(j|C_2)}$ holds. Therefore, $\frac{P(i|C_2)}{P(j|C_2)} = \frac{3}{2}$ holds. Set C_2 consists of only two alternatives, thus, $P(i|C_2) + P(j|C_2) = 1$. Simple algebra shows that $P(i|C_2) = \frac{3}{5}, P(i|C_2) = \frac{2}{5}$.

2. **[1.5 points]**

$$\Delta P(i) = \frac{3}{5} - \frac{1}{2} = \frac{1}{10}, \quad \Delta P(j) = \frac{2}{5} - \frac{1}{3} = \frac{1}{15},$$

therefore,

$$\frac{\Delta P(i)}{\Delta P(j)} = \frac{3}{2} = \frac{P(i|C_2)}{P(j|C_2)}.$$

If the two alternatives in the model exhibit IIA property, the probabilities for these alternatives change proportionally when the set of alternatives is changed.

3. **[1.5 points]** IIA property holds within, because the choice within the nest follows binary logit. Therefore, IIA holds for (i, k) and (j, l).

4. **[3 points]** The data suggests the following substitutions pattern: if an alternative from the set $\{\text{Auto}, \text{Walking}\}$ is removed, then the probabilities of choosing Bus or Rail increase in the same proportion. The same observation is valid for the set $\{\text{Bus}, \text{Rail}\}$ and choice of Auto and Walking. Therefore, IIA property holds (at least on the aggregate level) for these two sets. For other alternatives, such behavior does not hold. Therefore, the data suggests the structure of the nests as in Figure 2.

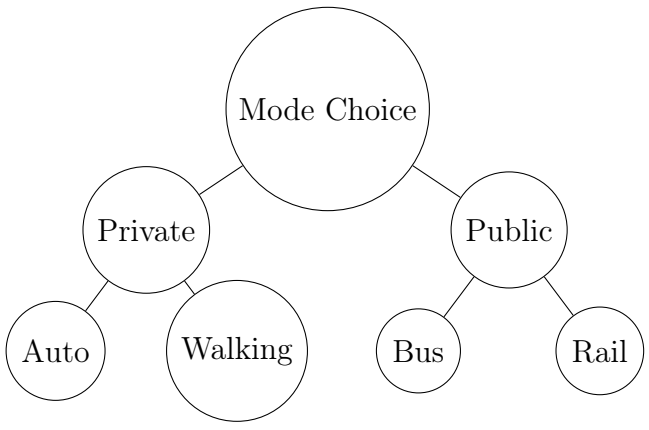


Figure 2: Mode Choice Tree Diagram

5. **[1 points]** There are several possible answers, but the simplest one is a follows. Use LL test with linear restriction $\mu_{m1} = \mu_{m2} = \mu$.

6. **[2.5 points]** To capture correlations with mixed logit we need to include the same error component into the alternatives within the same nests. For example, the following specification works

$$\begin{aligned} U_A &= V_A + \sigma_{Pr}\xi_{Pr} + \varepsilon_A; \\ U_W &= V_W + \sigma_{Pr}\xi_{Pr} + \varepsilon_W; \\ U_B &= V_B + \sigma_{Pb}\xi_{Pb} + \varepsilon_B; \\ U_R &= V_R + \sigma_{Pr}\xi_{Pb} + \varepsilon_R, \end{aligned}$$

where all error terms are independent from each other and $\varepsilon_i \sim \text{EV}(0,1)$. In principle, ξ_{Pr}, ξ_{Pb} may follow any valid distribution, for example, $\mathcal{N}(0,1)$.

To compare mixed logit with logit LL test with linear restriction $\sigma_{Pr}^2 = \sigma_{Pb}^2 = 0$ can be used.

Question 3

(20 points)

1. Switzerland is exploring the idea of raising energy prices to fund further renewable energy projects. As part of this initiative, they are conducting a study in a small municipality of 10,000 households. The focus is on understanding the number of people who have chosen to pay for clean energy in 2022, despite the additional cost of 20%. Households are divided into two segments according to their location: urban and rural. For every household, we know the average household income, the average number of residents, the percentage with high school education, and the percentage with access to renewable energy. A sample of 980 households was taken for the study in 2022.

Energy Source	Households		Total
	Urban	Rural	
Renewable	4000	1000	5000
Non-Renewable	3000	2000	5000
Total	7000	3000	10000

Table 7: Share of energy source choice in the population by household type in 2022

Energy Source	Households		Total
	Urban	Rural	
Renewable	500	100	600
Non-Renewable	300	100	400
Total	800	200	1000

Table 8: Share of energy source choice in the sample by household type in 2022

We have decided to use a binary logit model to study the choice of households for clean energy, focusing on the data from last year (2022). The estimates of the model are provided in Table 9.

Parameter	Estimate
ASC_R	-2.33
β_{cost}	-3.2
β_{income}	0.73
β_{rural}	-1.1
$\beta_{\text{household\_size}}$	0.4
$\beta_{\text{education}}$	1.2
$\beta_{\text{access\_renewable}}$	0.85

Table 9: Estimated parameters for the binary logit model

(a) **[6 points]** Write the specification of the model (utility functions) specifying which alternative will be the reference and do not forget to include all exogenous variables. After this, provide the diagram of the specification in which we can see the different explanatory variables, the utility, and the choice using the same drawing conventions as seen in the course. Do not forget to include n as a subindex for the individual, and i as a superindex for the choice when needed. In your discussion, specify which variables are endogenous and which are exogenous.

Now we extend our focus and try to use the available data from the 5 last years (2018-2022). Note that the households that we observe are the same over the years.

(b) **[0.5 points]** What are the benefits of introducing data from the last 5 years in our model? What can we take into account?
[1 point] Discuss potential correlation issues in the errors and in the past choices and their implications for modeling energy source choice.
[1.5 points] Draw a diagram to propose a specification that addresses one of these issues.

As a discrete choice modeling expert, the municipality has hired you and allocated a budget of 50,000 CHF to develop a survey. The goal is to refine the existing model. You decide to utilize this funding to add a new latent class to your model. This class will focus on understanding the awareness of households of environmental issues and their commitment to reducing environmental impact.

(c) **[3 points]** How would you call the latent class that will capture this phenomenon? Provide two specific question examples that would effectively serve as indicators for identifying this latent class?

(d) **[2 points]** Assume that a binary logit model was estimated and the coefficient for 'household income' in choosing renewable energy was found to be significant and positive (see Table 9). Interpret this result in the context of the choice of household energy sources.

To analyze the energy source choices of households, we apply a binary logit model. The model predicts the probability of a household choosing renewable energy over non-renewable energy, considering their urban or rural location. The estimated choice probabilities are given in Table 10.

(e) **[5 points]** Define the strata, and then calculate the weights of the sample from year 2022, and the estimated average market shares across rural and urban households.

Residence status	P(Renewable)	P(Non-Renewable)
Rural	0.8	0.2
Urban	0.4	0.6

Table 10: Choice probabilities of 2022 dataset

Assume the municipality implements a new policy to subsidize renewable energy for low-income households, proposing a 5% reduction in the renewable energy cost for these households.

(f) **[1 point]**: Theoretically, discuss how such a subsidy might influence the choice of renewable energy among different income groups, considering the existing model's structure. Discuss the potential implications for the model's key coefficients, especially how household income impacts energy costs.

CORRECTION

Question 3

(20 points)

1. (a) **[6 points]** In this context:

$$V_{Rn} = \text{ASC}_R + \beta_{\text{income}} \, \text{income}_n + \beta_{\text{cost}} \, \text{cost}_{Rn} + \beta_{\text{rural}} \, \text{rural}_n + \beta_{\text{size}} \, \text{size}_n + \beta_{\text{education}} \, \text{education}_n + \beta_{\text{access}} \, \text{access}_n + \varepsilon_{Rn},$$
$$V_{Un} = \beta_{\text{cost}} \, \text{cost}_{Un} + \varepsilon_{Un}.$$

   - The *endogenous variable* is the household's energy source choice [1 points].
   - *Exogenous variables* include cost of clean energy, household income, number of residents, level of education, the percentage with high school education, the percentage with access to renewable energy, and urban-rural classification [1 point].

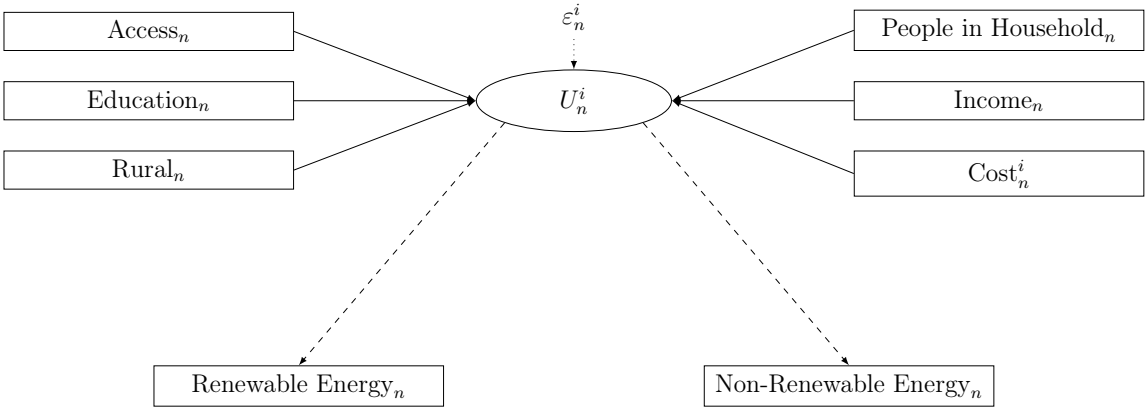
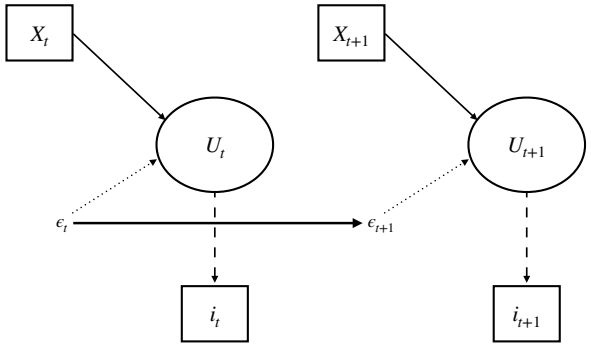
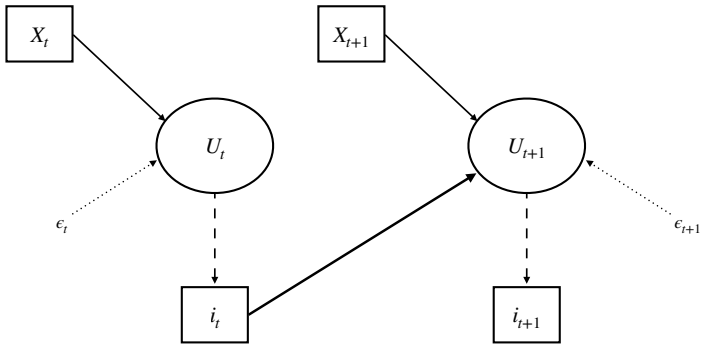


Figure 3: Choice model diagram for 2022

   (b) **[3 points]** Panel data allows to capture households choices over time, ensuring that previous choices and unobserved persistent household characteristics are appropriately factored into the analysis [0.5 points]. The key limitations of static models include serial correlation [0.5 point] and the fact that the choices might depend on the choices before the sampling period [0.5 point].

   (c) **[3 points]** The phenomenon to capture would be the environmentally friendly latent class (Note that any answer that makes sense is accepted).





This involves identifying individuals who demonstrate a strong inclination towards environmental conservation, sustainable living, and eco-friendly practices.

(d) **[2 points]** A positive coefficient for household income in a binary logit model suggests higher-income households are more likely to choose renewable energy, possibly due to financial capabilities or environmental consciousness. This indicates the importance of income in determining renewable energy choice.

(e)
- **[5 point]**

$$\hat{W}(j) = \frac{1}{S} \sum_{g=1}^{4} S_g \omega_g P_g(j) \text{ [1 point]}$$

- If we define the strata as g=1 as Urban and Renewable, g=2 as Rural and Renewable, g=3 as Urban and Non-Renewable, and g=4 as Rural and Non-Renewable [1 point] :

– [0.25 points] $\omega_1 = \frac{N_1}{S_1} \cdot \frac{S}{N} = \frac{4000}{500} \cdot \frac{1000}{10000} = \frac{4}{5}$

- [0.25 points] $\omega_2 = \frac{1000}{100} \cdot \frac{1000}{10000} = 1$
- [0.25 points] $\omega_3 = \frac{3000}{300} \cdot \frac{1000}{10000} = 1$
- [0.25 points] $\omega_4 = \frac{2000}{100} \cdot \frac{1000}{10000} = 2$

• [1 point]

$$\begin{aligned}\hat{W}(\text{Renewable}) &= \frac{1}{1000}\left(500 \cdot \omega_1 \cdot 0.4 + 100 \cdot \omega_2 \cdot 0.8 + 300 \cdot \omega_3 \cdot 0.4 + 100 \cdot \omega_4 \cdot 0.8\right) \\ &= 0.76\end{aligned}$$

or $\hat{W}(\text{Renewable}) = 1 - \hat{W}(\text{Non-Renewable})$

• [1 point]

$$\begin{aligned}\hat{W}(\text{Non-Renewable}) = \frac{1}{1000}&\left(500 \cdot \omega_1 \cdot 0.6 + 100 \cdot \omega_2 \cdot 0.2\right. \\ &\left. + 300 \cdot \omega_3 \cdot 0.6 + 100 \cdot \omega_4 \cdot 0.2\right) \\ &= 0.24\end{aligned}$$

or $\hat{W}(\text{Non-Renewable}) = 1 - \hat{W}(\text{Renewable})$

(f) **[1 point]**

- Income and Energy Cost Interaction: The subsidy would theoretically make renewable energy more affordable for low-income households. This affordability could increase the propensity of these households to choose renewable energy, altering the relationship between income and energy choice. In your model, this would suggest a stronger positive correlation between lower-income levels and the choice of renewable energy.
- Other Variables: The subsidy might also interact with other variables like household size or location. For instance, if larger households or rural households have lower incomes on average, the subsidy might disproportionately benefit these groups.

Question 4

(20 points)

Part 1 [16 Points].

A researcher collected data using a stated preference (SP) survey where each individual had to choose between:

- Alternative 1: car
- Alternative 2: bus
- Alternative 3: metro

All alternatives had **travel time** (in minutes) and **travel cost** (in CHF) as attributes. Moreover, bus and metro had **headway time** (in minutes) as attribute. Each respondent faced T different choice tasks and all alternatives were always available. In total, N individuals completed the survey.

Moreover, the respondents answered to four attitudinal questions. The researcher decided to estimate an Integrated Choice and Latent Variable (ICLV) model using the attitudinal questions as indicators of latent variables.

1. **[2 points]** The initial step taken with the indicators involves conducting a factor analysis. Explain the nature of this analysis and the type of information it can yield to aid in specifying the model.
2. **[1 points]** What does a latent variable represent? Conceptually, what is its causal relationship with the indicators?
3. **[2 points]** The researcher correctly decided to incorporate the attitudinal questions as indicators of latent variables. Please explain the two reasons why the indicators cannot be used directly as explanatory variables in the utility function.

After some initial analysis, the researcher concluded in a model specification with two latent variables specified as:

- Latent variable 1 (LV1):
  - $I_{1,1}$: 5-point Likert scale, ordered indicator
  - $I_{1,2}$: 5-point Likert scale, ordered indicator
- Latent variable 2 (LV2):
  - $I_{2,1}$: 4-point Likert scale, ordered indicator

- $I_{2,2}$: 4-point Likert scale, ordered indicator

Both latent variables were specified as a linear function of **income** and **age**.

4. **[2 points]** Provide a generic figure that presents the model specification, following the description provided (consider all attributes provided in the description of the data collection). Use the same drawing conventions as seen in the course.

5. **[3 points]** For each variable (latent or observed), mention if it is an output of the model, an attribute or an individual characteristic.

6. Based on the model specification that you provided in your diagram:

   - **[3 points]** Write the utility functions related to each of the modes.
   - **[2 points]** Write the structural equations of the latent variables.
   - **[1 point]** Pick one of the indicators and write the measurement equation.

   For your answer, consider the following:

   - Assume mode choice is based on a logit model.
   - Assume the utility of mode choice is linear in parameters.
   - In the equations, only include parameters that need to be estimated (for instance, the scale μ of the utility functions is fixed to 1, hence it does not need to be estimated). Do not fix/normalise more parameters than the minimum required.
   - Assume alternative specific parameters.
   - Include all the error terms, together with the associated parameters and distributional assumptions.
   - The error terms of the latent variables are normally distributed.

Part 2 [4 Points].

A researcher conducted a survey on bike ownership and then estimated an ICLV model using maximum simulated likelihood. The conceptual framework of the model is presented in Figure 4.

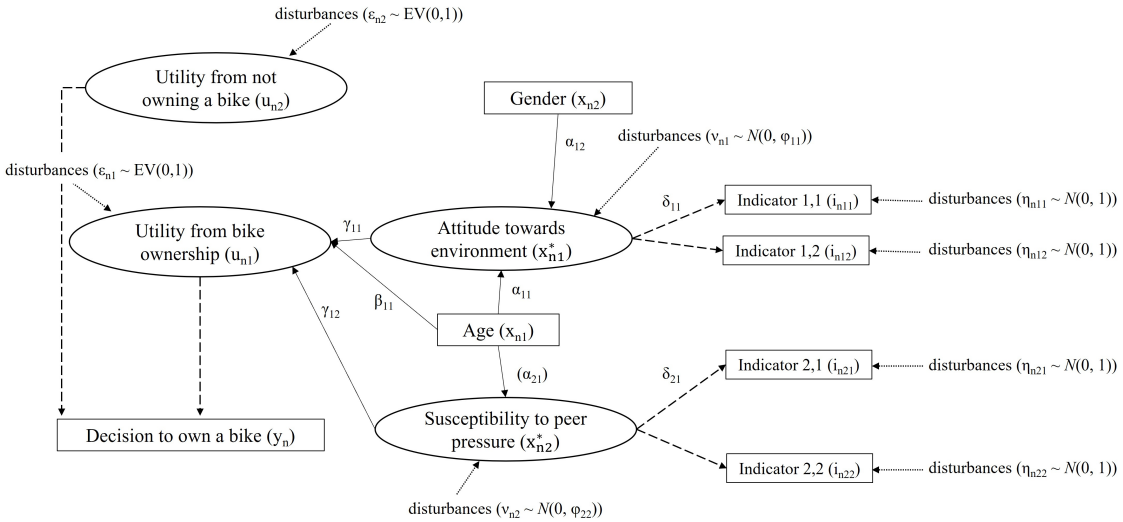


Figure 4: The ICLV bike ownership model framework (Source (adaptation from): Vij & Walker, 2016)

The researcher then investigated how this ICLV model predicted choice probabilities compared to an error-component logit mixture model, which was a reduced form of the ICLV (Figure 5) [*both models also had an ASC parameter estimated which is not shown in the figures*]:

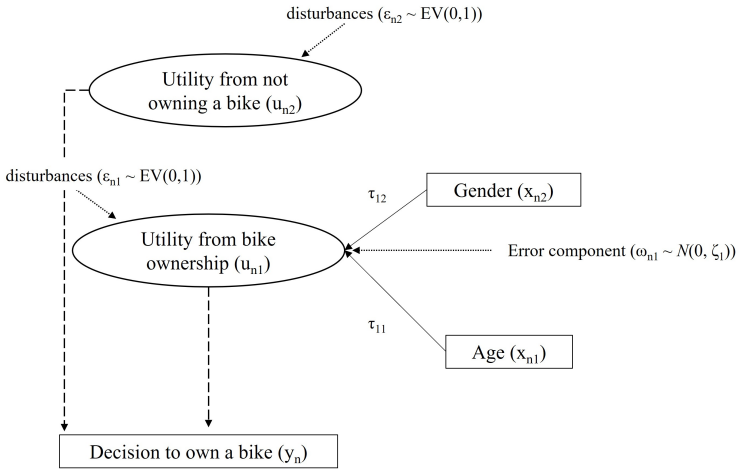


Figure 5: The error components reduced logit mixture

1. **[2 points]** Suppose that we do not have access to the indicators. In this case, the two models have equivalent specifications. Can you explain why? *Hint: Write the utility functions for both models.*

2. **[2 points]** If we use the indicators, explain why the log-likelihood functions of the two models are different.

CORRECTION

Question 4

(20 points)

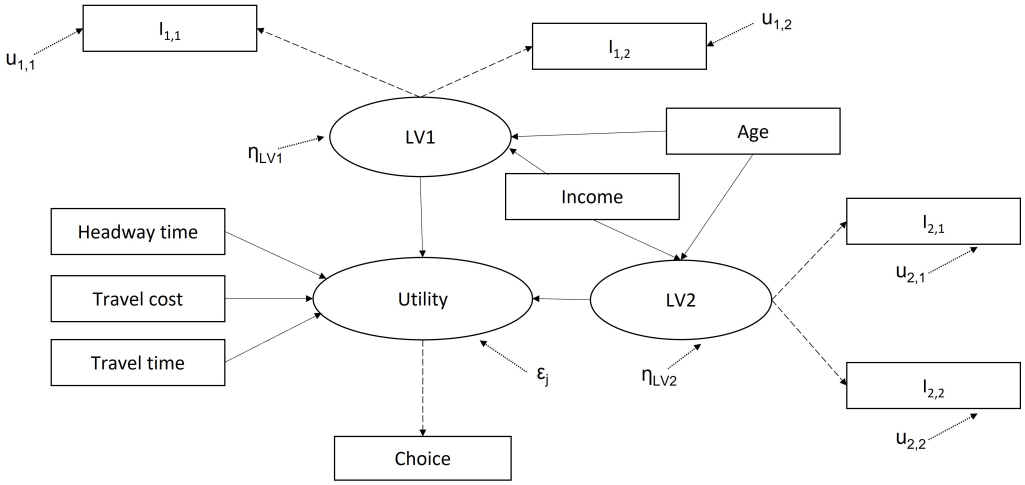
Part 1

1. **[2 points]** Factor analysis is used to identify underlying variables, or factors, from a set of observed variables. The analysis returns as results a number of coefficients (factor loadings) related to each factor that suggest the correlation of every observed variable to the factor. Based on the magnitude of each loading, the researcher decides if the respective item is associated to each factor. The relationship between items and factors is decided based on the sign of the loading. Ultimately, each factor is treated as a candidate latent variable for the ICLV model.

2. **[1 point]** A latent variable represents a variable that is not directly observed. However, we can "measure" the impact of the latent variables on the psychometric indicators. In that sense, in our model specification we assume that a latent variable affects the value of an indicator and the latter is treated as a dependent variable.

3. Indicators cannot be used as explanatory variables due to (a) measurement errors (arbitrary scale, interpretation of the scale, justification bias, overreaction in responses) and (b) no forecasting possibility (there is no way to predict indicators in the future)

4. **[2 points]** The model figure is:



5. The variables are:

- **[1 point]** The outputs of the model are the choice and the four indicators.
- **[1 point]** The attributes of the alternatives are the travel time, cost, headway and the utility functions.
- **[1 point]** The characteristics of an individual n are the two latent variables, gender, and income.

6. The question has multiple answers depending on whether the student decides to fix the variance of the latent variables or their impact to one of the indicators. Both solutions are correct and accepted.

   The following normalisations need to be applied for identification (assuming the scales of all alternatives $\mu_j = 1$):

   - For each parameter related to the choice probabilities, we can estimate $J-1$ cases, where J is the number of alternatives. The simplest approach is to fix all parameters of the car utility to 0.
   - The standard deviation (consequently variance) of the latent variables is fixed to 1.
   - The standard deviation (consequently variance) of the latent variables can be estimated, if the effect to one of the indicators is fixed to 1.

   The fully specified model is:

   [3 points] *Utility functions*:

   $U_{car,nt} = V_{car,nt} + \epsilon_{car,nt} = ASC_{car} + \beta_{time,car} * time_{car,nt} + \beta_{cost,car} * cost_{car,nt} + \lambda_{LV1,car} * LV1_n + \lambda_{LV2,car} * LV2_n + \epsilon_{car,nt}$

   $U_{bus,nt} = V_{bus,nt} + \epsilon_{bus,nt} = ASC_{bus} + \beta_{time,bus} * time_{bus,nt} + \beta_{cost,bus} * cost_{bus,nt} + \beta_{headway,bus} * headway_{bus,nt} + \lambda_{LV1,bus} * LV1_n + \lambda_{LV2,bus} * LV2_n + \epsilon_{bus,nt}$

   $U_{metro,nt} = V_{metro,nt} + \epsilon_{metro,nt} = ASC_{metro} + \beta_{time,metro} * time_{metro,nt} + \beta_{cost,metro} * cost_{metro,nt} + \beta_{headway,metro} * headway_{metro,nt} + \lambda_{LV1,metro} * LV1_n + \lambda_{LV2,metro} * LV2_n + \epsilon_{metro,nt}$

   where (explanations below are optional as not explicitly requested in the question):

   - ASC_{car}, ASC_{bus}, ASC_{metro}: are the constants of car, bus, and metro respectively

- $\beta_{time,car}$, $\beta_{time,bus}$, $\beta_{time,metro}$: are the travel time parameters of car, bus, and metro respectively
- $\beta_{cost,car}$, $\beta_{cost,bus}$, $\beta_{cost,metro}$: are the travel cost parameters of car, bus, and metro respectively
- $\beta_{headway,bus}$, $\beta_{headway,metro}$: are the time headway parameters of bus and metro respectively
- $\lambda_{LV1,car}$, $\lambda_{LV1,bus}$, $\lambda_{LV1,metro}$: are the LV1 parameters of car, bus, and metro respectively (impact of LV1 on utilities)
- $\lambda_{LV2,car}$, $\lambda_{LV2,bus}$, $\lambda_{LV2,metro}$: are the LV2 parameters of car, bus, and metro respectively (impact of LV2 on utilities)
- $\epsilon_{car,nt}$, $\epsilon_{bus,nt}$, $\epsilon_{metro,nt}$: are EV(0,1) i.i.d error terms of the car, bus, and metro utilities respectively, for the t^{th} observation of the n^{th} individual

Comments:

- If the student proposes generic parameters in the utilities (not alternative specific), the answer is considered wrong.
- If parameters are alternative specific but there are mistakes in normalisations (e.g. estimate all parameters in all utilities), remove the point from one of the utilities.

[2 points] *Latent variables*:

$$LV1 = \gamma_{Income_{LV1}} * (Income) + \gamma_{Age_{LV1}} * (Age) + \eta_{LV1,n}$$
$$LV2 = \gamma_{Income_{LV2}} * (Income) + \gamma_{Age_{LV2}} * (Age) + \eta_{LV2,n}$$

where:

- $\gamma_{Age_{LV1}}$, $\gamma_{Age_{LV2}}$ parameters associated to the impact of age on the latent variables.
- $\gamma_{Income_{LV1}}$, $\gamma_{Income_{LV2}}$ parameters associated to the impact of income on the latent variables.
- $\eta_{LV1,n}$ is a random error $\sim N(0, \sigma^2_{LV1})$ across n.
- $\eta_{LV2,n}$ is a random error $\sim N(0, \sigma^2_{LV2})$ across n.

The fully specified model may also have constant parameters. These are not identified and "absorbed" by the ASC of the choice model.

[1 point] *Measurement equations*:

For the indicators we choose $I_{1,1}$ and $I_{2,1}$. The answers for $I_{1,2}$ or $I_{2,2}$ have the same structure respectively.

$$I_{1,1} = \left\{ \begin{array}{l} 1 \text{ if } I^*_{1,1,n} < \tau_{11,1} \\ 2 \text{ if } \tau_{11,1} \leq I^*_{1,1,n} < \tau_{11,2} \\ 3 \text{ if } \tau_{11,2} \leq I^*_{1,1,n} < \tau_{11,3} \\ 4 \text{ if } \tau_{11,3} \leq I^*_{1,1,n} < \tau_{11,4} \\ 5 \text{ if } \tau_{11,4} \leq I^*_{1,1,n} \end{array} \right\}$$

where:

$$I_{1,1,n} = \delta_{I_{1,1}} + \zeta_{I_{1,1}} * LV1_n + \upsilon_{1,1,n},\ \upsilon_{1,1} \sim N(0, \sigma^2_{1,1})$$
$$\tau_{11,1} = -\kappa_{11,1} - \kappa_{11,2}$$
$$\tau_{11,2} = -\kappa_{11,1}$$
$$\tau_{11,3} = \kappa_{11,1}$$
$$\tau_{11,4} = \kappa_{11,1} + \kappa_{11,2}$$

$$I_{2,1} = \left\{ \begin{array}{l} 1 \text{ if } I^*_{2,1,n} < \tau_{21,1} \\ 2 \text{ if } \tau_{21,1} \leq I^*_{2,1,n} < \tau_{21,2} \\ 3 \text{ if } \tau_{21,2} \leq I^*_{2,1,n} < \tau_{21,3} \\ 4 \text{ if } \tau_{21,3} \leq I^*_{2,1,n} \end{array} \right\}$$

where:

$$I_{2,1,n} = \delta_{I_{2,1}} + \zeta_{I_{2,1}} * LV2_n + \upsilon_{2,1,n},\ \upsilon_{2,1} \sim N(0, \sigma^2_{2,1})$$
$$\tau_{21,1} = \kappa_{21,1} - \kappa_{21,2}$$
$$\tau_{21,2} = \kappa_{21,1}$$
$$\tau_{21,3} = \kappa_{21,1} + \kappa_{21,2}$$

Part 2

1. **[2 points]** The two choice models have equivalent specifications. To understand why the ICLV model results in a specification which is a reduced form mixed logit model, we need to work on the specification of the former and reach to the specification of the latter. This is easier to derive if we use the simulated log-likelihood formula as follows (where $y_{nj} = 1$ if the respondent owns a bike)[1]:

$$P_y = \frac{1}{Q}\sum_{1=1}^{Q}\prod_{j=1}^{J}\left[\frac{exp(\beta_{j*}\chi_n + \gamma_j\chi^*_{nq})}{\sum_{j'=1}^{J} exp(\beta_{j'*}\chi_n + \gamma_{j'*}\chi^*_{nq})}\right]^{y_{nj}} =$$

$$\frac{1}{Q}\sum_{1=1}^{Q}\prod_{j=1}^{J}\left[\frac{exp(\beta_{j*}\chi_n + \gamma_{j*}(A\chi_n + \nu_{nq}))}{\sum_{j'=1}^{J} exp(\beta_{j'*}\chi_n + \gamma_{j'*}(A\chi_n + \nu_{nq}))}\right]^{y_{nj}} =$$

$$\frac{1}{Q}\sum_{1=1}^{Q}\prod_{j=1}^{J}\left[\frac{exp((\beta_{j*} + \gamma_{j*}A)\chi_n + \gamma_{j*}\nu_{nq})}{\sum_{j'=1}^{J} exp((\beta_{j'*} + \gamma_{j'*}A)\chi_n + \gamma_{j'*}\nu_{nq})}\right]^{y_{nj}}$$

where ν_{nq} is the q^{th} draw from $N(0, \Phi)$. The term χ^*_{nq} is a vector of explanatory variables for the latent variable while χ_n is the vector of explanatory variables of the utilities related to the choices. The terms $\beta_{j*}, \gamma_{j*}, A$ are all parameters to be estimated. Hence, they can all be replaced by:

$\tau_{j'*} = \beta_{j'*} + \gamma_{j'*}A$

Similarly, we can define $\omega_n = \gamma_{j'*}\nu_{nq}$ as a vector of random variables normally distributed such that ω_{nq} is the q^{th} draw from $N(0, Z)$. Keeping this changes in mind we redefine the formula as:

$$\frac{1}{Q}\sum_{1=1}^{Q}\prod_{j=1}^{J}\left[\frac{exp(\tau_{j'*}\chi_n + \omega_n))}{\sum_{j'=1}^{J} exp(\tau_{j'*}\chi_n + \omega_n)}\right]^{y_{nj}}$$

This solution collapses to the solution of an error component mixed logit model (without latent variables) with a utility specification:

$$u_n = Tx_n + \omega_n + \epsilon_n$$

[1]The problem can be also solved using the ICLV utility function solely (the student does not need to suggest the solution with the random draws, but only develop the utility of the ICLV model or suggest the propsoed solution in words without using formulae). This is still correct and accepted.

2. **[2 points]** The observations associated with each individual are different. In Figure 4, both the choice and the indicators. In Figure 5, only the choice.