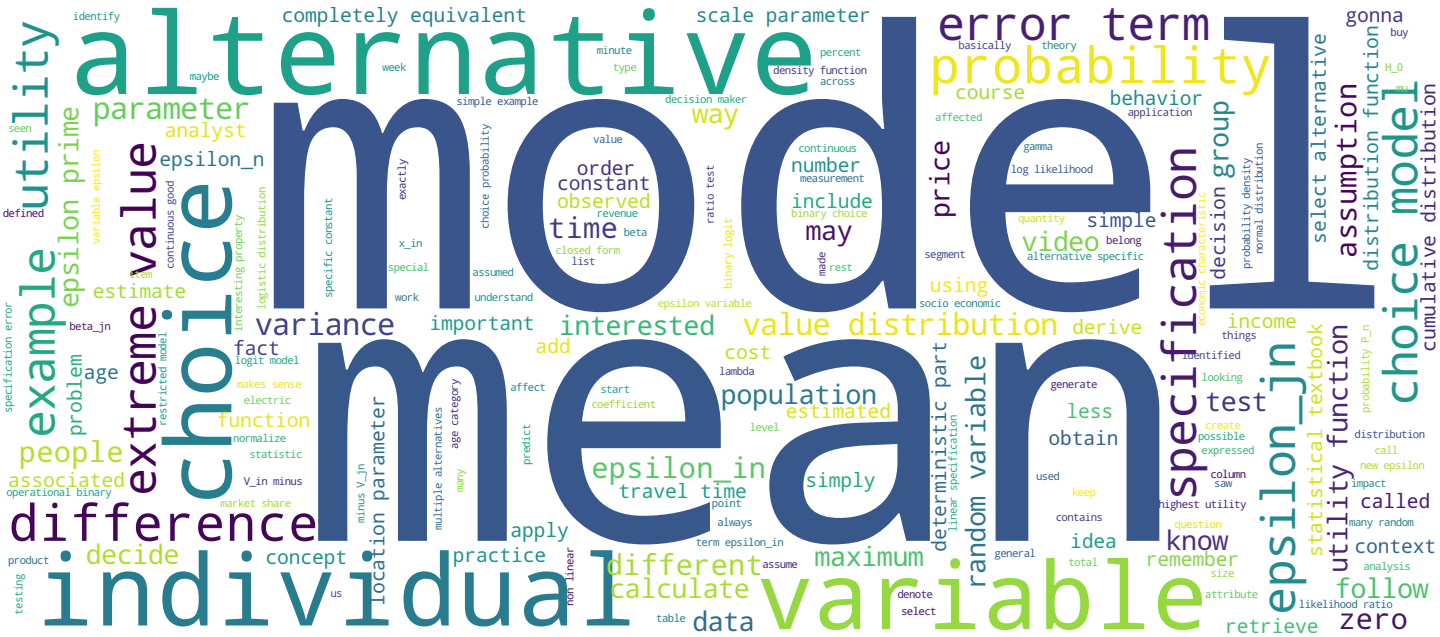
Binary choice

Model specification: the error term

Virginie Lurkin

Introduction to choice models

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE



Search MOOC

Video

Binary choice model

Hello everyone, and welcome to this video. In the previous videos, you have been introduced to the concept of the individual decision-maker, who faced with a choice set of feasible discrete alternatives, is going to select the one with the highest utility. For many different reasons, we have explained you that this utility is best viewed as a random variable. In today's video, we gonna be interested in the special case where the choice set contains only two alternatives. You may think about alternative i as the decision to take the car to go to work, and alternative j as the decision to take the train, for example. And the goal of this video will be to derive an operational binary choice model. Why binary? Because the choice set contains only two alternatives. You may be wondering why we start with a choice set with only two alternatives. The reason is very simple. It's that it is a very simple case. So that we can introduce many ideas and concepts very easily. But many of those ideas and concepts will remain valid when we will extend the analysis to choice sets containing multiple alternatives.

Notes

Summary

0m 04s

Binary choice model

Two alternatives: $\mathcal{C}_n = \{i, j\}$

$$\begin{aligned} U_{in} &= V_{in} + \varepsilon_{in} \\ U_{jn} &= V_{jn} + \varepsilon_{jn} \end{aligned}$$

Choice model

$$\begin{aligned} P_n(i|\{i,j\}) &= \Pr(U_{in} \geq U_{jn}) \\ &= \Pr(V_{in} + \varepsilon_{in} \geq V_{jn} + \varepsilon_{jn}) \\ &= \Pr(V_{in} - V_{jn} \geq \varepsilon_{jn} - \varepsilon_{in}) \\ &= \Pr(\varepsilon_n \leq V_{in} - V_{jn}) \end{aligned}$$

where $\varepsilon_n = \varepsilon_{jn} - \varepsilon_{in}$.

So today, we are interested in special case where the choice set C_n contains only two alternatives: i and j. Each alternative has a utility U_in and U_jn. And this utility is composed of two main terms: the first term V_in, V_jn, is the deterministic part, and represent everything that can be observed by the analyst. While the second term, epsilon_in, epsilon_jn, is everything that cannot be observed by the analyst, plus any measurement or specification errors. The choice model is then the probability P_n(i) that individual n is going to select alternative i. And it's the probability that this alternative i creates a greater utility than j. Substituting the utility expression above in this expression, I have this. And if I rearrange the terms, I have this. Finally, I decide to call epsilon_n the difference between epsilon_jn and epsilon_in I have that the probability for individual n to choose alternative i is simply the probability that epsilon_n is less or equal than V_in minus V_jn. What is very important to notice here is that only the difference of deterministic utilities does matter for the choice set.

Notes

Summary

1m 13s

Error term

Three assumptions about the random variables ε_{in} and ε_{jn}

1. What's their mean?
2. What's their variance?
3. What's their distribution?

Note

- For binary choice, it would be sufficient to make assumptions about $\varepsilon_n = \varepsilon_{jn} - \varepsilon_{in}$.
- But we want to generalize later on.

An important component of any operational choice model is the error term. And now, we gonna see what are the assumptions that are made about those random variables epsilon_in and epsilon_jn. And we gonna try to answer three questions: first is what's their mean, then what's their variance, and finally what's their distribution. We note here that, for the binary situation, it will be be sufficient to make assumptions about the difference between the error terms, epsilon_n. But because we would like to generalize later on, and we would like to create a story that is easily extended to the more general case of multiple alternatives within a choice set, we gonna do assumptions on each epsilon variable, epsilon_in and epsilon_jn.

Notes

Summary

2m 41s

The mean

Change of variables

- Define $\mathsf{E}[\varepsilon_{in}] = \beta_{in}$ and $\mathsf{E}[\varepsilon_{jn}] = \beta_{jn}$.
- Define $\varepsilon'_{in} = \varepsilon_{in} - \beta_{in}$ and $\varepsilon'_{jn} = \varepsilon_{jn} - \beta_{jn}$,
- so that $\mathsf{E}[\varepsilon'_{in}] = \mathsf{E}[\varepsilon'_{jn}] = 0$.

Choice model

$$
\begin{aligned}
P_n(i|\{i,j\}) = & \\
& \Pr(V_{in} - V_{jn} \geq \varepsilon_{jn} - \varepsilon_{in}) = \\
& \Pr(V_{in} - V_{jn} \geq \varepsilon'_{jn} + \beta_{jn} - (\varepsilon'_{in} + \beta_{in})) = \\
& \Pr(V_{in} + \beta_{in} - (V_{jn} + \beta_{jn}) \geq \varepsilon'_{jn} - \varepsilon'_{in}) = \\
& \Pr(V_{in} + \beta_{in} - (V_{jn} + \beta_{jn}) \geq \varepsilon'_{n})
\end{aligned}
$$

where $\varepsilon'_n = \varepsilon'_{jn} - \varepsilon'_{in}$.

So, the mean... Let's assume that the mean of epsilon_in is beta_in and the mean of epsilon_jn is beta_jn. It could be any value. Then, an interesting property is that I can always define a new epsilon variable, epsilon prime, as the difference between the previous error term, epsilon_in minus the mean of epsilon_in, beta_in. Doing so, I can obtain a new epsilon variable, epsilon prime in, with a mean equals to zero. Doing so does not affect the choice probability, as you can see here. This probability is completely equivalent to this one, which is also equivalent to this one, and it's completely equivalent to this one. And again, we use the same trick. We define a new epsilon prime n variable, which is the different between epsilon prime jn and epsilon prime in.

Notes

Summary

3m 33s

Alternative specific constant

- The mean of each error term can be moved to the deterministic part.
- It is captured by a parameter, to be estimated from data.
- It is called the alternative specific constant (ASC).
- Only the mean of the difference of the error terms is identified.

So, in definitive, what we can see is that the mean of each error term can be moved to the deterministic part. And this mean will be captured by a parameter and will be estimated from the data. This mean is called the "alternative specific constant". And the very important thing to remember is that only the mean of the difference of the error terms, so epsilon_n, is identified.

Notes

Summary

4m 32s

Shift invariance

The choice model is not affected by a uniform shift of all utility functions

$$P_n(i|\{i,j\}) = \Pr(U_{in} \geq U_{jn}) = \Pr(U_{in} + K \geq U_{jn} + K) \quad \forall K.$$

An interesting property of the choice model is that, the choice model is not affected by a uniform shift of all utility functions. So it means that I can decide to add or to remove a constant K from the utility of each alternative, so i and j in this case, and this does not affect the choice probability.

Notes

Summary

5m 00s

Alternative Specific Constant

Many equivalent specifications

$$
\begin{array}{rcll}
U_{in} & = & V_{in} + \beta_{in} & +\varepsilon'_{in} \\
U_{jn} & = & V_{jn} + \beta_{jn} & +\varepsilon'_{jn} \\
 & & \text{or} & \\
U_{in} & = & V_{in} & +\varepsilon'_{in} \\
U_{jn} & = & V_{jn} + \beta_{jn} - \beta_{in} & +\varepsilon'_{jn} \\
 & & \text{or} & \\
U_{in} & = & V_{in} + \beta_{in} - \beta_{jn} & +\varepsilon'_{in} \\
U_{jn} & = & V_{jn} & +\varepsilon'_{jn}
\end{array}
$$

In practice

Normalize one constant to zero and estimate $\beta_n = \beta_{jn} - \beta_{in}$

So, as shown before, it means that the utility in and utility jn can be expressed as follows: so we removed the mean of the epsilon error term variables from the random term to the deterministic parts. But this specification is also completely equivalent to this one, which consists to add the difference between the means only to the utility of the j variable. And it's also completely equivalent to this one, which is to add the difference between the means to the utility i. All those specifications are completely equivalent. And, in practice, what is generally done, is to normalize one constant to zero and to estimate beta_n, that is the difference between beta_jn and beta_in.

Notes

Summary

5m 24s

Scale invariance

The choice model is not affected by a uniform scaling of all utility functions

$$P_n(i|\{i,j\}) = \Pr(U_{in} \geq U_{jn}) = \Pr(\alpha U_{in} \geq \alpha U_{jn}) \quad \forall \alpha > 0.$$

Another interesting property is that the choice model is not affected by a uniform scaling of all utility functions. And so, what you can see here, is that I can decide to multiply each utility by a constant alpha without affecting the choice model.

Notes

Summary

6m 16s

The variance

$$\begin{aligned} \mathrm{Var}(\alpha U_{in}) &= \alpha^2 \mathrm{Var}(U_{in}) \\ \mathrm{Var}(\alpha U_{jn}) &= \alpha^2 \mathrm{Var}(U_{jn}) \end{aligned}$$

The variance is not identified

- As any α can be selected arbitrarily, any variance can be assumed.
- No way to identify the variance of the error terms from data.
- The scale has to be arbitrarily decided.

In practice

The scale parameter of the assumed distribution is normalized to 1.

So what about the variance? As you can see here, the issue is that the variance is not identified. Any alpha can be selected arbitrarily, and so any variance can be assumed. It also means that there is no way to identify the variance of the error terms from data. So, the scale has to be arbitrarily decided. In practice, what is done, for convenience, is to normalize the scale parameter to one.

Notes

Summary

6m 33s

The distribution

Assumption

ε_{in} and ε_{jn} are the maximum of many r.v. capturing unobservable attributes (e.g. mood, experience), measurement and specification errors.

Gumbel theorem

The maximum of many i.i.d. random variables approximately follows an Extreme Value distribution: EV(η, μ), with $\mu > 0$.

Finally, what's the assumption regarding the distribution of those error terms? The assumption made here is that epsilon_in and epsilon_jn are the maximum of many random variables, that capture everything that cannot be observed by the analyst. So, it can be things that I cannot observe, such as mood, experience, but it's also measurement and specification errors. Note that, assuming that epsilon_in and epsilon_jn are the maximum of many random variables is somehow consistent with the maximum utility theory, since we are interested in the alternative that is going to generate the highest utility, it makes sense to be interested in the maximum of all those unobservable attributes. From statistics, we know that the maximum of many identically and independently distributed random variables approximately follows an extreme value distribution, with location parameter eta and scale parameter mu.

Notes

Summary

7m 00s

The Extreme Value distribution EV(η, μ)

Probability density function (pdf)

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

Cumulative distribution function (CDF)

$$\begin{aligned} P(c \geq \varepsilon) = F(c) &= \int_{-\infty}^{c} f(t)dt \\ &= e^{-e^{-\mu(c-\eta)}} \end{aligned}$$

Then, from statistical textbooks, we can retrieve the probability density function of the extreme value distribution, which is expressed as follows, and we can also retrieve the cumulative distribution function (CDF), which takes this form. What is very interesting here, is that this cumulative distribution function takes a closed form. This is a very nice property that's gonna be very useful later.

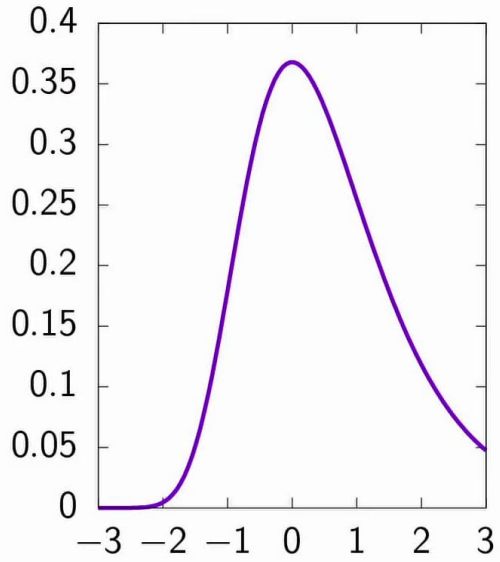
Notes

Summary

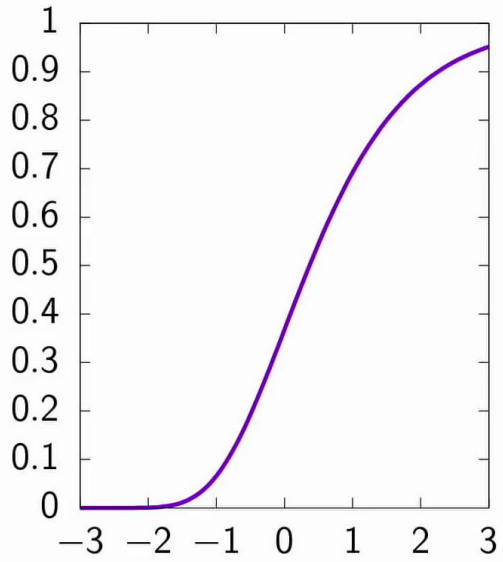
8m 05s

The Extreme Value distribution

pdf EV(0,1)



CDF EV(0,1)



So here, you can see the probability density function of the extreme value distribution, and you can see that, unlike the normal distribution, the distribution is not symmetric. So, you can see that it is askew, and it has a bigger right tail. And it makes sense as the probability of higher values is higher, as we take the maximum of those values. And here, you can see the cumulative distribution function.

Notes

Summary

8m 33s

The Extreme Value distribution

Properties

If

$$\varepsilon \sim \mathsf{EV}(\eta, \mu)$$

then

$$\mathsf{E}[\varepsilon] = \eta + \frac{\gamma}{\mu} \quad \text{and} \quad \mathsf{Var}[\varepsilon] = \frac{\pi^2}{6\mu^2}$$

where γ is Euler's constant.

Euler's constant

$$\gamma = -\int_0^\infty e^{-x} \ln x \, dx \approx 0.5772$$

Again, from statistical textbooks, we can retrieve the properties of the extreme value distribution. So, here are the mean and the variance of epsilon, if epsilon follows an extreme value distribution. And, gamma is simply the Euler constant.

Notes

Summary

9m 01s

The distribution

Assumptions

- ε_{in} and ε_{jn} are i.i.d. Extreme Value.
- If an alternative specific constant is in the model, their mean can be assumed to be any constant.
- It is convenient to set the scale parameter to 0, so that $\mathrm{E}[\varepsilon_{in}] = \mathrm{E}[\varepsilon_{jn}] = \gamma/\mu$.

Distributions

$$\varepsilon_{in} \sim EV(0, \mu), \qquad \varepsilon_{jn} \sim EV(0, \mu)$$

Problem

We need the distribution of $\varepsilon_n = \varepsilon_{jn} - \varepsilon_{in}$

So, now, what are the assumptions on epsilon_in and epsilon_jn? It is that epsilon_in and epsilon_jn are following an i.i.d. extreme value distribution. And, remember that, if we include an alternative specific constant in our model, then it means that the mean can be assumed to be any constant. And, in practice, it is convenient to set the location parameter to zero, so that the mean of the error terms is gamma over mu. So now, we have the distribution of epsilon_in and the distribution of epsilon_jn. They both follow an extreme value distribution with location parameter equal to zero, and mu as a scale parameter. The problem is that, what we need, is the distribution of epsilon_n, that is the difference between epsilon_jn and epsilon_in.

Notes

Summary

9m 20s

Logistic distribution

From the properties of the extreme value distribution, we have

$$\begin{aligned} \varepsilon_{in} &\sim \text{EV}(0, \mu) \\ \varepsilon_{jn} &\sim \text{EV}(0, \mu) \\ \varepsilon_n = \varepsilon_{jn} - \varepsilon_{in} &\sim \text{Logistic}(0, \mu) \end{aligned}$$

Unlike the normal distribution, the difference between two extreme value distributions does not follow an extreme value distribution. But, again from statistical textbooks, we know that the difference between two extreme value distributed variables follows a logistic distribution. And the location parameter is zero, and the scale parameter is mu.

Notes

Summary

10m 09s

The Logistic distribution: Logistic(η,μ)

Probability density function (pdf)

$$f(t) = \frac{\mu e^{-\mu(t-\eta)}}{(1 + e^{-\mu(t-\eta)})^2}$$

Cumulative distribution function (CDF)

$$P(c \geq \varepsilon) = F(c) = \int_{-\infty}^{c} f(t)dt = \frac{1}{1 + e^{-\mu(c-\eta)}}$$

with $\mu > 0$.

Again, if we go and check our statistical textbook, we will retrieve the probability density function of the logistic and, again, the very good news for us is that this cumulative distribution function has a closed form. as you can see here.

Notes

Summary

10m 34s

The binary logit model

Choice model

$$P_n(i|\{i,j\}) = \Pr(\varepsilon_n \leq V_{in} - V_{jn}) = F_\varepsilon(V_{in} - V_{jn})$$

The binary logit model

$$P_n(i|\{i,j\}) = \frac{1}{1 + e^{-\mu(V_{in} - V_{jn})}} = \frac{e^{\mu V_{in}}}{e^{\mu V_{in}} + e^{\mu V_{jn}}}$$

And so, finally, it becomes easy to derive the operational binary choice model. So, the probability P_n(i) that individual n is going to select alternative i, is the probability that this epsilon_n is less or equal than V_in minus V_jn. So, now that we know that epsilon_n follows a logistic distribution, and that we know the cumulative distribution function for epsilon n, it becomes trivial to compute the binary logit model, that is the probability P_n(i) that individual n is going to select alternative i. And, as you can see, it is a closed form expression, that is a function of V_in and V_jn.

Notes

Summary

10m 49s

I hope that you have enjoyed the video, and that you have understood how to derive an operational binary logit model. In the rest of the week, you're gonna learn how to estimate those models. But, first, I invite you to test yourself by doing the exercises that are proposed.

Notes

Summary

11m 32s