Introduction to Identification

Abi Adams

HT 2017

Outline for Today

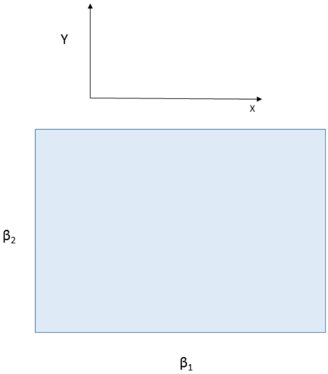
Aim: Understand basic concepts so that we can move on to apply them in a range of applied settings in future lectures

I Recap

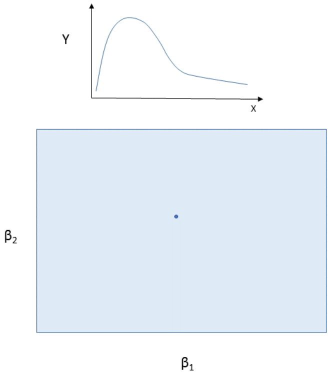
II Normalisations (in the context of discrete choice)

III Special Regressors

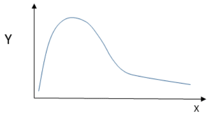
Recap

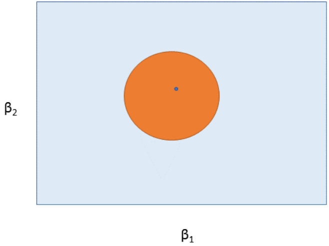


Recap

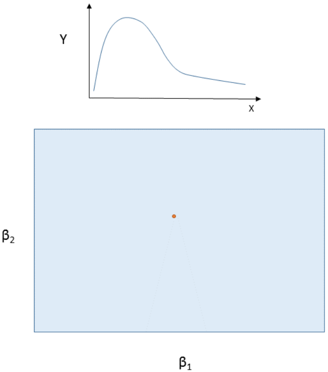


Recap: Observational Equivalence

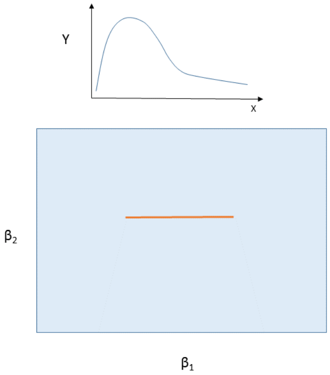




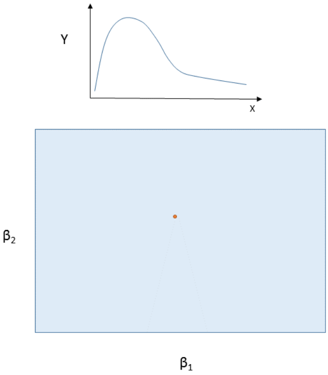
Recap: Point Identification



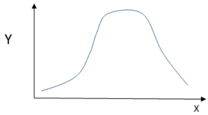
Recap: Structural Features

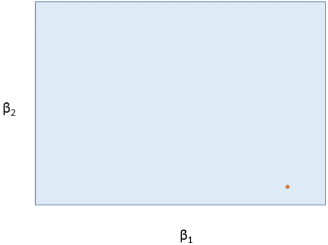


Recap: Uniform Identification



Recap: Uniform Identification





Recap: Terminology

- **Observational Equivalence 1**
  S' and S'' such that $F_{YX}^{S'} = F_{YX}^{S''}$ are observationally equivalent
- **Observational Equivalence 2**
  S' and S'' such that $F_{\phi}^{S'} = F_{\phi}^{S''}$ are observationally equivalent given the features of the data that are knowable, ϕ
- The model Γ **identifies** S^0 if there is no $S' \in \mathcal{M}_\Gamma$ such that $F_{\phi}^{S^0} = F_{\phi}^{S'}$

Proving Identification

- There are a number of ways that one might demonstrate identification
- The most common way is to prove identification **by construction**: given a structure, one is able to write a closed form expression for θ as a function of ϕ
- However, not necessary that a closed form expression exists for a structure to be identified

Example: Identifying WTP

- Lewbel, Linton, McFadden (2011): want to recover the distribution of people's willingness to pay (WTP), $W^\star$, $F_{W^\star}(w)$.
- Dataset used by Hanemann et al. (1991) to elicit the WTP for protecting wetland habitats and wildlife in California's San Joaquin Valley
- For each person in the sample, researchers draw a price P from a known distribution function and ask if they would be willing to pay $\$P$ or more to preserve the wetland

Example: Identifying WTP

- Binary choice: D denotes an individual's response

$$D = I\left(W^{\star} > P\right) \tag{1}$$

- Given random assignment of prices, P is distributed independently of $W^{\star}$

$$\begin{aligned} E(D|P = p) &= Pr\left(W^{\star} > P|P = p\right) \\ &= Pr\left(W^{\star} > P\right) \\ &= 1 - Pr\left(W^{\star} < P\right) \\ &= 1 - F_{W^{\star}}(p) \end{aligned} \tag{2}$$

- Here, identification is proved by construction: $F_{W^{\star}}$ is uniquely determined by the function $E(D|P = p)$, which is assumed to be known given ϕ

Example: Identifying WTP

- Note that given the experimental design, the function $F_{W^\star}$ might not be identified everywhere
- In the motivating experiment, P could take on one of 14 values between \$25 and \$375 — can identify the distribution function only at $w^\star = p$ at these particular values
- To identify the entire distribution function $F_{W^\star}$, would want to design an experiment so that P could take on any value that $W^\star$ could equal — p should be drawn from a continuous distribution with support at least as large as the range of possible values of $W^\star$

Proving Identification: Extremum

- Another common method is to prove that θ^0 is the unique solution to some maximisation problem defined given S
- E.g. Show that the likelihood is globally concave, then maximum likelihood will have a unique maximising value
  - Establish identification by showing that the unique maximiser in the population equals the true θ^0
- Note trend to attempt to show this with complicated structural models by graphing marginal likelihood function at the estimated parameter vector — don't do this for presentations!

Discrete Choice

- The WTP Example provides some insight into special regressor methods, often used in discrete choice models when we want to be flexible about the distribution of unobserved preference heterogeneity
- Before starting, a brief recap on discrete choice to allow us to discuss the role of normalisations

Binary Choice

- Imagine a consumer choosing whether to consume a good/enter into treatment/start working
- Choose the action if:

$$\alpha + \beta X_i > \epsilon_i \tag{3}$$

- The probability that they choose

$$\begin{aligned} Pr(Y_i = 1 | X = x_i) &= Pr(\alpha + \beta x_i > \epsilon_i | X = x_i) \\ &= Pr(\alpha + \beta x_i > \epsilon_i) \\ &= F_\epsilon(\alpha + \beta x_i) \end{aligned} \tag{4}$$

Binary Choice

- In most of the models you will have encountered thus far, you proceed by putting a functional formal assumption on the distribution of the errors
- Example: Probit: $\epsilon_i \sim N\left(\mu, \sigma^2\right)$
- However, without further restrictions, $\{\alpha, \beta\}$ are **not identified**

$$\begin{aligned} Pr(Y_i = 1 | X = x_i) &= F_\epsilon(\alpha + \beta x_i) \\ &= \Phi\left(\frac{\alpha + \beta x_i - \mu}{\sigma}\right) \end{aligned} \quad (5)$$

Binary Choice: Location

- Different combinations of $\{\mu, \alpha\}$ are observationally equivalent

$$\begin{aligned} Pr(Y_i = 1|X = x_i) &= \Phi\left(\frac{\alpha + \beta x_i - \mu}{\sigma}\right) \\ &= \Phi\left(\frac{(\alpha + \kappa) + \beta x_i - (\mu + \kappa)}{\sigma}\right) \\ &= \Phi\left(\frac{\widetilde{\alpha} + \beta x_i - \widetilde{\mu}}{\sigma}\right) \end{aligned} \tag{6}$$

- Standard: restrict $\mu = 0$ — the **location normalisation**

$$Pr(Y_i = 1|X = x_i) = \Phi\left(\frac{\alpha + \beta x_i}{\sigma}\right) \tag{7}$$

Binary Choice: Scale

- Different combinations of $\{\sigma, \alpha, \beta\}$ are observationally equivalent

$$
\begin{aligned}
Pr(Y_i = 1|X = x_i) &= \Phi\left(\frac{\alpha + \beta x_i}{\sigma}\right) \\
&= \Phi\left(\frac{\kappa\alpha + \kappa\beta x_i}{\kappa\sigma}\right) \\
&= \Phi\left(\frac{\widetilde{\alpha} + \widetilde{\beta} x_i}{\widetilde{\sigma}}\right)
\end{aligned} \tag{8}
$$

- Standard: restrict $\sigma = 1$ — the **scale normalisation**

$$
Pr(Y_i = 1|X = x_i) = \Phi(\alpha + \beta x_i) \tag{9}
$$

Normalisations

- In parametric models, common to impose these restrictions on the distribution of the error term as we have just seen
- For example, in the Probit model above, assume that ϵ has a standard normal distribution
- However, note that we could have imposed the location and scale restrictions on $\{\alpha, \beta\}$ rather than $\{\mu, \sigma\}$
  - For example, $\alpha = 0$ and $\beta^k = 1$, allowing ϵ to have an arbitrary mean and variance

Normalisations

- Normalisations common in semi- and nonparametric models
- Example: structure is a linear index model

$$E(Y|X) = g(\alpha + X\beta) \tag{10}$$

- Features of interest: $\theta = \{g, \beta, \alpha\}$
- Normalisations/restrictions typically imposed on parameter vectors in semiparametric models

Normalisations

- For any nonzero constant κ, define $\widetilde{\theta} = \{\widetilde{g}, \widetilde{\beta}, \widetilde{\alpha}\}$ with $\widetilde{\beta} = \beta/\kappa$, $\widetilde{\alpha} = \alpha/\kappa$ and $\widetilde{g}(z) = g(\kappa z)$
  - Then $\widetilde{\theta}$ is observationally equivalent to θ
  - All elements $\widetilde{\beta}$ in the identified set have $\widetilde{\beta}$ proportional to β — identified up to a scale
  - Require a scale normalisation, usually $\beta^k = 1$ or $\beta'\beta = 1$

Normalisations

- For any nonzero constant κ, define $\widetilde{\theta} = \{\widetilde{g}, \beta, \widetilde{\alpha}\}$ with $\widetilde{\alpha} = \alpha + \kappa$ and $\widetilde{g}(z) = g(z - \kappa)$
  - Then $\widetilde{\theta}$ is observationally equivalent to θ
  - Require a location normalisation, usually $\alpha = 0$ — exclude a constant

Normalisations

- What makes something a normalisation rather than a restriction?
- Calling a restriction a normalisation implies that is does not restrict or limit behaviour — 'without loss of generality'
- Thus, whether a restriction can be thought of in this way depends in part on how we will use and interpret the model

Normalisations

- If one is simply interested in, e.g. marginal effects, then the scale normalisation is indeed without loss of generality

$$\frac{\partial Pr(Y_i = 1)}{\partial X} = \frac{\beta}{\sigma}\phi\left(\frac{\alpha + \beta X - \mu}{\sigma}\right) \tag{11}$$

- If however want to imbue coefficients with meaning, one might need to be careful!
- Caution: direct comparison of discrete choice coefficients across different samples/specifications

Normalisations: Outside Options

- ‘Outside option’ normalisations are also common in discrete choice models
- Let utility from choice $Y = y$ for $y = 0, 1$

$$\alpha_y + \beta_y X + \epsilon_y \tag{12}$$

- Utility maximisation means that choose good 1 if:

$$\begin{aligned} \alpha_1 + \beta_1 X + \epsilon_1 &> \alpha_0 + \beta_0 X + \epsilon_0 \\ (\alpha_1 - \alpha_0) + (\beta_1 - \beta_0) X + (\epsilon_1 - \epsilon_0) &> 0 \\ \alpha + \beta X + \epsilon &> 0 \end{aligned} \tag{13}$$

- Interpret $\alpha + \beta X$ as the utility from $Y = 1$ if assume the normalisation that the utility of the outside option is zero

$$\alpha_0 + \beta_0 X + \epsilon_0 = 0 \tag{14}$$

Normalisations: Outside Options

- In static discrete choice models, this is usually a free normalisation, without loss of generality
- However, this might not be the case in dynamic discrete choice models
- Assuming that the outside option has the same utility in every period imposes real restrictions on preferences and hence on behaviour — be careful!

Relaxing Assumptions on ϵ

- The parametric assumptions placed on the distribution of unobserved errors are essentially arbitrary and can have very restrictive behavioural implications
- To introduce some more common concepts in the literature, explore the use of 'special regressors' in discrete choice and their role in identification
- Intuitively, variation in special regressors allow one to trace out the distribution of unobservables

Special Regressor Methods

- Let's pick up on the example from the beginning of the lecture, casting in the standard notation used in the literature:

$$\begin{aligned} D &= I\left(W^{\star} > P\right) \\ &= I\left(W^{\star} - P > 0\right) \\ &= I\left(W^{\star} + V > 0\right) \end{aligned} \tag{15}$$

- Let $H(v) = E(D|V = v)$ and suppose V is continuously distributed (V is the special regressor!)

$$\begin{aligned} H(v) &= Pr\left(W^{\star} + V > 0\right) \\ &= Pr\left(W^{\star} > -V\right) \\ &= 1 - F_{W^{\star}}(-v) \end{aligned} \tag{16}$$

- If the support of V contains the support of $-W^{\star}$, then the entire distribution function $F_{W^{\star}}$ would be identified

Special Regressor Methods

- Want to identify, e.g., the average willingness to pay; the special regressor allows one to do this

$$
\begin{aligned}
E(W^\star) &= \int_{w_l}^{w_u} w f_{w^\star}(w)\,\mathrm{d}w \\
&= \int_{w_l}^{w_u} w \frac{\partial F_{w^\star}(w)}{\partial w}\,\mathrm{d}w \\
&= \int_{w_l}^{w_u} w \frac{\partial\,[1 - H(-w)]}{\partial w}\,\mathrm{d}w
\end{aligned} \tag{17}
$$

Special Regressor Methods

- Key assumptions on the special regressor:
  - Independence (or conditional independence)
  - Additive
  - Continuity
  - Large support
- Pop up a lot even if not always identified as special regressors!

Special Regressor Methods

- Large support important for identification of certain features that rely on knowledge of the tails of a distribution

$$E(W^\star) = \int_{w_l}^{w_u} w \frac{\partial\,[1 - H(-w)]}{\partial w}\,\mathrm{d}w \qquad (18)$$

- If supp(V) bounded to $a \leq V \leq b$, then $F_{W^\star}(w)$ only identified for $-b \leq W^\star \leq -a$
- In this case, $E(W^\star)$ is not even set identified
- No bounds on $E(W^\star)$ because $F_{W^\star}$ could have mass arbitrarily far below $-b$ or above $-a$

Special Regressor: Random Coefficients

- Random coefficients often used to allow for more sophisticated unobserved preference heterogeneity specifications

$$Y = I(V\epsilon_V + X\epsilon_X > 0) \tag{19}$$

where ϵ_V and ϵ_X are random coefficients

- Assume $\epsilon_V > 0$ and let $\epsilon = \epsilon_X/\epsilon_V$ — a scale normalisation

$$Y = I(V + X\epsilon > 0) \tag{20}$$

- Is the distribution of ϵ identified from the data?

Special Regressor: Random Coefficients

- Assume that V is a special regressor, distributed independently of X

$$\begin{aligned} Y &= I(V + X\epsilon > 0) \\ &= I(V + U > 0) \end{aligned} \tag{21}$$

- Using the same argument as before $F_{U|X}$ is identified by variation in the special regressor

$$\begin{aligned} E(Y|X = x, V = v) &= Pr(v + U > 0|X = x) \\ &= 1 - F_{U|X}(-v) \end{aligned} \tag{22}$$

- So the distribution of ϵ is identified!

Conclusion

- These two lectures have introduced the concept of observational equivalence and introduced its role in proving identification of structural features
- Next week we will consider the connection between this ‘structural’ approach to identification and a ‘causal’ approach to identification
- We will apply these results to consider identification in simple equilibrium settings