Econometrics Notes

Priliminary and very incomplete

T. Lamadon (lamadon@uchicago.edu)

October 20, 2020

Contents

1 Probabilities **4**
- 1.1 Independence . . . 5
- 1.2 Random variables . . . 5
- 1.3 Change of variables, function of a random variable . . . 7
- 1.4 Notions of convergence . . . 8
- 1.5 Characteristic Functions for random variables . . . 8
- 1.6 The law of large numbers . . . 9
- 1.7 The central limit theorem . . . 9

2 Models, parameters and objects of interest **9**
- 2.1 Identification . . . 10
- 2.2 Samples and Inference . . . 11

3 Ordinary Least Squares **12**
- 3.1 Non-iid samples - clustered standard errors . . . 13

4 Parametric Inference **13**
- 4.1 Maximum likelihood . . . 13
- 4.2 Discrete choices and McFadden Multinomial choice . . . 16
- 4.3 Dynamic discrete choice . . . 17
- 4.4 Moment based estimation . . . 17

5 Random effect models **17**
- 5.1 A running example: unemployment hazard rate . . . 18
- 5.2 Estimation using EM . . . 19
- 5.3 On the identification of Hazard rate models . . . 21
- 5.4 Random effect versus fixed effect . . . 21
- 5.5 Dynamic discrete choice with unobserved types . . . 21

6 Linear regression with many regressors **22**
- 6.1 Stepwise selection . . . 23
- 6.2 Ridge Regression . . . 23
- 6.3 Lasso . . . 23
- 6.4 Principal component analysis . . . 23

7 Non parametric regression **24**
- 7.1 Choice of tuning parameter . . . 25
- 7.2 Kernel estimator . . . 25
- 7.3 Sieve Estimator . . . 25
- 7.4 Regression tree . . . 26
- 7.5 Semi-parametric estimator . . . 26

8 Boostrap **26**

9 Fixed effect model **28**
9.1 Marginal effect . . . 29
9.2 Group-fixed effect estimators . . . 30

10 Linear Panel **30**
10.1 Omitted variable . . . 30
10.2 Measurement error in regressors . . . 31
10.3 Autoregressive model . . . 31
10.4 Example, Firm money demand . . . 32

11 Recap of important concepts: **32**

A Additional notes **33**
A.1 Integration . . . 33
A.2 o and o_p notations . . . 33
A.3 KL divergence . . . 34
A.4 Linear algebra refresher . . . 34
A.5 Bellman principle of optimality . . . 34

1 Probabilities

A probability space is a triplet $(\Omega, \mathcal{F}, P)$ where:

- **Sample space Ω:** set of all possible outcomes $\omega \in \Omega$.
- **Set of events $\mathcal{F}$**, any $A \in \mathcal{F}$ is called an event, we have that $A \subseteq \Omega$ with following properties:
  - $\emptyset \in \mathcal{F}$
  - $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
  - $A_1, A_2, ... \in \mathcal{F} \Rightarrow \cup_i A_i \in \mathcal{F}$
- **Probability measure** $P : \mathcal{F} \to \mathbb{R}$ with following properties:
  1. $P(A) \geq 0$ for all $A \in \mathcal{F}$
  2. $P(\Omega) = 1$
  3. If $A_1, A_2, ...$ possibly infinite, are disjoints ($i \neq j \Rightarrow A_i \cap A_j = \emptyset$) then $P(\cup_i A_i) = \sum_i P(A_i)$

From there we can derive several properties:

- $A \subseteq B \Rightarrow P(A) \leq P(B)$
- $P(A \cap B) \leq \min\{P(A), P(B)\}$

Two simple examples

Discrete. Think of the tossing a six-sided die, then $\Omega = \{1, 2, 3, 4, 6\}$, the most general $\mathcal{F}$ is the power set of Ω (all possible combinations) and P is simply defined in the case of a fair die as the length of the event since each individual outcome can be attributed a given mass probability.

Continuous. Think of spinning a bottle where the angle is the outcome. Again let's think of a fair spin. In this case $\Omega = [0, 2\pi]$. Now $\mathcal{F}$ and P are more interesting here. We can't construct this probability measure from the probability of each point since we would want that for any $\omega \in \Omega$, the probability of such event should be 0! What we would want is for $\mathcal{F}, P$ to be compatible with a simple notion of mass such that any interval $[a, b]$ is in $\mathcal{F}$ and$P([a, b]) = a - b$. One can achieve this in this framework by defining $\mathcal{F}$ as the set of all open sets in $[0, 2\pi]$ and define P exactly such that for any segment $[a, b)$ we have that $P([a, b)) = a - b$. This σ-algebra generated by all open intervals is called the Borel measure (or the Lesbegue measure when all ω singletons are also included).

Using power set for $\mathcal{F}$ - Vitali sets This could be a tempting idea, however in the case where Ω is a continuum, the power set includes sets that can't be assigned a measure without generating contradictions. The Vitali sets are such sets. Intuitively, the Vitali sets split the $[0,1]$ line into an infinite but countable group of sets. These sets turn are also translation of each other, which means that if Lesbegue measurable, they will have the same measure. Now, if this measure is 0 then $[0,1]$ would have measure 0, which is not true. If the measure of each set if $\epsilon > 0$, by countable reunion, $[0,1]$ would have measure ∞ which is also not true. Hence they can't be measurable, and can't be included in $\mathcal{F}$.

1.1 Independence

- Two events $A, B \in \mathcal{F}$ are independent iif $P(A \cap B) = P(A) \cdot P(B)$.
- Two σ-algebras $\mathcal{F}_1$ and $\mathcal{F}_2$ are independent iif $P(A \cap B) = P(A) \cdot P(B) \quad \forall A \in \mathcal{F}_1, B \in \mathcal{F}_2$

1.2 Random variables

Definition 1. A real random variable X is a function $X : \Omega \to \mathbb{R}$ such that $\forall r \in \mathbb{R}$, $\{\omega : X(\omega) \leq r\} \in \mathcal{F}$.

In this case we say that X is measurable with respect to $\mathcal{F}$. If the condition is not satisfied it is as if X would provide a "thiner" cut of omega in the sense that some event would have ambiguous values. This becomes clearer when we define the expectation.

Note. This can be generalized to many sets beyond $\mathbb{R}$. In particular we can think of multidimensional random variables. TBD.

Distribution function for random variables

For a random variable X on $(\Omega, \mathcal{F}, P)$ we can define the CDF function as:

$$F_X(x) = P(\{X \leq x\})$$

where the event $(X \leq x) \in \mathcal{F}$ for sure since X is a random variable (see definition). In the case where $F_X(x)$ is differentiable, one can introduce the PDF that for $A \in \mathcal{F}$ we have:

$$Pr[X \in A] = \int_A f \mathrm{d}P = \int_{\omega \in A} X(\omega) P(\mathrm{d}\omega)$$

where the integral is defined on level sets. In many cases it is equivalent to the Riemann integral which we know best as:

$$Pr[X \in A] = \int_x f(x) \mathrm{d}P(x)$$

and gives

$$F_X(x) = \int_0^x f_X(u) du$$

We can then define independence between random variables.

- Two r.v. X, Y are independent if $F_{X,Y}(x,y) = F_x(x) \cdot F_y(y)$

Expectations, moments, conditional expectation

We also define the expectation operator:

$$\mathbb{E}(X) = \int X \mathrm{d}P$$

which in the case of discrete reduces to summing over the different events:

$$\mathbb{E}(X) = \sum_{\omega \in \Omega} X(\omega) P(\omega)$$

and in the continuous case means summing over the level sets of X

$$\mathbb{E}(X) = \int X(\omega) P(\mathrm{d}\omega)$$

we can then introduce more generally moments, such that for any measurable function g we have :

$$\mathbb{E}(g(X)) = \int g(X(\omega)) P(\mathrm{d}\omega)$$

and note in particular that

$$E(1[X \leq x]) = F_X(x)$$

Conditional Expectations

We commonly want to consider the realization of events conditional on another event having realized. For a sub-event set $\mathcal{F}_2$ (which needs to satisfy the properties listed at the beginning, meaning being a σ-algebra on Ω) we can define $\mathbb{E}(X|\mathcal{F}_2)$ as the unique function that satisfies:

$$\int_A E(X|\mathcal{F}_2) \mathrm{d}P = \int_A X \mathrm{d}P \qquad \forall A \in \mathcal{F}_2$$

which using the indicator function can help us define the condition CDF for any random variable:

$$F_{X|Y=y}(x) = \int_A E(1[X \leq x]|\mathcal{F}_2) \mathrm{d}P$$

- todo: Optimality properties of conditional expectations
- todo: Conditioning as factorization

Expectation, Variance, Covariance

Definitions

$$Var(X) = \mathbb{E}(X - \mathbb{E}(X))^2$$
$$Cov(X, Y) = \mathbb{E}(X - \mathbb{E}(X))(Y - \mathbb{E}(Y)$$

we have the following properties:

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$$
$$Var[X] = \mathbb{E}Var[X|Y] + Var[\mathbb{E}[X|Y]]$$

and of course:

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$
$$Var(X + Y) = Var(X) + Var(Y) + 2cov(X, Y)$$

1.3 Change of variables, function of a random variable

Let's first look at a one dimensional random variable, We consider the random variable $Y = g(X)$ for g increasing and seek its CDF and PDFs

We start from the definitions:

$$F_X(x) = Pr[X \leq x] = Pr[g(X) \leq g(x)] = Pr[Y \leq g(x)]$$
$$f_X(x) = g'(x) f_Y(g(x))$$

In the context of many variables we get that:

$$f_Y(y) = \frac{1}{g'(g^{-1}(y))} f_X(g^{-1}(x))$$

For the multivariate case, consider $Y = r(X)$, then

$$f_Y(\mathbf{y}) = f_X(\mathbf{x}) \left| det \left(\frac{dx}{dy} \right) \right|$$

where

$$\left(\frac{dx}{dy} \right)_{ij} = \frac{\partial x_i}{\partial y_j}$$

Let's look at a simple example. Consider indexing points on a disk, and think of the uniform distribution.

$$f_{XY}(x, y) = \frac{1}{\pi} \quad \forall x^2 + y^2 \leq 1$$

and next we want to consider the polar coordinates. This is the following transformation:

$$x = r \cdot \cos(\theta)$$
$$y = r \cdot \sin(\theta)$$

we then compute

$$\frac{dx}{dy} = \begin{bmatrix} \cos\theta & -r \cdot \sin\theta \\ \sin\theta & r \cdot \cos\theta \end{bmatrix}$$

which gives a determinant of $r\cos\theta^2 + r\sin\theta^2 = r$ and so we get that

$$f_{r\theta}(r,\theta) = \frac{r}{2\pi}$$

- examples: The paradox of the three jewelry boxes
- link: http://www.math.uah.edu/stat/dist/Transformations.html#cov4

Manifestation of the Borel-Kolmogorov paradox Indeed, here notice that when conditioning on $\theta = 0$ the radius r is equal to x when we condition on $y = 0$. However, we see that $f_{r\theta}$ is proportional to r whereas f_{xy} is just $1/\pi$. Why is it not a paradox? Density functions are defined with respect to a measure, and when conducting a change of variable and setting $y = 0$ in one case and $\theta = 0$ in the other, we end up with 2 different conditional measures. In other words, how we narrowed to the set $\{(x,y)$ st. $y = 0\}$ matters. This problem arises only when one conditions on a set of measure 0. Note that it remains true that for any subset A of the circle we still have that $\int_A f_{r\theta}(r,\theta)\mathrm{d}r\mathrm{d}\theta = \int_A f_{xy}(x,y)\mathrm{d}x\mathrm{d}y$, even when $A = \{(x,y)$ s.t. $\theta = 0\}$ since we get 0 on both sides. Another way to think about it is to ask the question how was the set $\{(x,y)$ st. $y = 0\}$ picked, was it picked by randomly picking y or was it chosen by randomly picking θ. Each approach delivers a different conditional density.

1.4 Notions of convergence

We introduce 3 notions of convergence among random variables:

- Convergence in distribution $X_n \xrightarrow{d} X$ iif $P(X_n \leq x) \to P(X \leq x)$ for all x
- Convergence in probability $X_n \xrightarrow{p} X$ iif $P(||X_n - X|| > \epsilon) \to 0$ for all ϵ
- Convergence almost surely $X_n \xrightarrow{a.s.} X$ iif $P(\lim_{n\to\infty} ||X_n - X|| = 0) = 1$ for all ϵ

Example "An example is if your random variables are just the characteristic functions of intervals determined by the angles [n, n+1/n). The area of non-convergence to zero is 1/n which goes to zero in length, but each given point on the circle will be 1 infinitely many times, so the set of points where convergence does not happen has measure 1."

1.5 Characteristic Functions for random variables

An important transformation of random variables is the characteristic function. For a given random variable X with k dimensions is it given by

$$t \mapsto \mathbb{E}e^{it^\intercal X}, \quad t \in \mathbb{R}^k$$

properties:

- $X_n \xrightarrow{d} X$ iif $\mathbb{E}e^{it^\intercal X_n} \to \mathbb{E}e^{it^\intercal X}$ point-wise in t (Levy's continuity Theorem)
- A normal distribution $\mathcal{N}(\mu, \Sigma)$ has characteristic function $e^{it^\intercal \mu - \frac{1}{2}t^\intercal \Sigma t}$
- $\mathbb{E}X^n = \frac{\partial^n}{\partial t^n}\phi_x(\frac{t}{i}) = \mathbb{E}e^{t^\intercal X}$ is the moment generating function (se example for the normal distribution)
- $\phi_x(t) = 1 + it\mathbb{E}X + \frac{(it)^2}{2!}\mathbb{E}X^2 + \frac{(it)^3}{3!}\mathbb{E}X^3$.....

1.6 The law of large numbers

we now use the characteristic functions to establish an important result, the fact that $\frac{1}{n}\sum_i X_i \xrightarrow{p} \mu$

$$\begin{aligned}\mathbb{E}e^{it\bar{X}_n} &= \left(\mathbb{E}e^{i\frac{t}{n}X}\right)^n \\ &= \left(1 + i\frac{t}{n}\mu + o(\frac{1}{n})\right)^n \to e^{it\mu}\end{aligned}$$

We get that the series of random variable formed of the average converges to the true mean of these random variables.

1.7 The central limit theorem

perhaps more surprisingly is how this sequence converges to in distribution towards this constant: (assumes iid, mean is 0 and variance$\mathbb{E}X^2$)

$$\mathbb{E}e^{it\sqrt{n}\bar{X}_n} = \left(\mathbb{E}e^{i\frac{t}{\sqrt{n}}X}\right)^n = \left(1 + 0 - \frac{1}{2}\frac{t^2}{n}\mathbb{E}X^2 + o(\frac{1}{n})\right)^n \to e^{-\frac{1}{2}t^2\mathbb{E}X^2}$$

2 Models, parameters and objects of interest

Now that we have clear concepts of probability we move to our main goal which is to learn about models using data. We start by defining a **population** representation of our data which we denote in general as $F(Y, X)$. From this population object we will consider **samples**. A sample is a sequence of random variables with some connection to the population object $F(Y, X)$. Here we will focus on random samples (and panel samples, see later), meaning that we have n independent draws $(Y_i, X_i)_{i=1..n}$ from $F(Y, X)$. The next step is to develop a model for our data $F(X, Y)$. A model will simply be a class of data generating processes indexed by a parameter θ (potentially of infinite dimension). In general we can denote a model as $M = \{F(Y, X; \theta), \theta \in \Theta\}$. It will be quite common for a model to introduce variables are not directly observed in the data (or population data). For instance it might introduce a residual with statistical properties. Going forward, we will continuously make the distinction between observable and non-observable variables.

Example 2. A linear *model* for $F(Y, X)$ can be defined as $M = \{Y = \beta X + \epsilon; (\beta, F_{X\epsilon})\}$.

We see that such a model might restrict strongly the set of possible generated data. In addition in introduces the variable ϵ. Using this example, consider a restricted class that does not assumes additive error $M = \{Y = f(X, \epsilon); (f, F_{x\epsilon})\}$. Here the parameters are the joint distribution $F_{\epsilon,X}(\epsilon, x)$ and some function f_0. There will be two main set of task that econometricians/statisticians will be interested in:

1. **In support prediction**: This is concerned with quantifying events that only rely on conditioning on observable quantities. For example given a generative model such as $M = \{Y = f(X, \epsilon); (f, F_{x\epsilon})\}$ and a parameter guess $(f^{(1)}, F^{(1)}_{x\epsilon})$ the researcher would like to get an as good as possible prediction for Y_i given the value X_i. The researcher could care about the following quantity:

$$||Y - f(X, 0)||$$

2. **Causal effect & out of support predictions**: This is concerned with quantifying events that might rely on conditioning on variables that are unobserved OR with quantifying effects for combinations of observable variables that are not seen in the data.

There is an extremely popular model that clarifies notion of causal effect, the potential outcome model. Consider an outcome Y and a binary treatment$T$. The population is then given by $F(Y, T)$. Next consider a model where $Y_i = \alpha_i + \beta_i T_i + \epsilon_i$. We actually allow for the effect of the treatment to be heterogenous. A prediction question could be to ask what is the average value of Y_i for each of the treatment values $T = 0, 1$. It is relatively easy in the this context as one could simply look at the realized distribution. Often prediction questions will become difficult when the conditioning set becomes very large. A causal question however would be to ask what is $\mathbb{E}[\beta_i | T = 1]$ which is the average treatment effect on the treated.

Application: class size Let's consider an actual economic question that has been studied. Researchers have long wondered about the effect of class sizes on children performance.

2.1 Identification

The first notion that we want to introduce is the notion of identification. It addresses the question of whether we can hope to recover the parameters of a Model from observed data.

Definition 3. For a given $M = \{F(Y, X; \theta), \theta \in \Theta\}$ we say that θ is identified with respect to $F(Y, X)$ iif there a unique θ^* such that $F(Y, X; \theta^*) \stackrel{d}{=} F(Y, X)$.

This has a simple corollary which states that if two parameters generate observationally equivalent distribution, then the model is not identified. Two important points to note:

1. identification is a statement about bout the model class jointly with the data $F(Y, X)$

2. identification is an argument about the population object $F(Y, X)$ not about a finite sample coming from it.

Going further, people have developed theory about **partially identified model**. For instance it might not be impossible to pin point exactly f from the data in an particular model class, but it might be possible to pin down it's sign. In general we can think about a feature $g(\theta)$ of a model as being identified w.r.t. $F(Y, X)$.

2.2 Samples and Inference

Of course in general we do not have the population object $F(Y, X)$, we have a sample $(Y_i, X_i)_{i=1..n}$. We are then interested in learning about the parameters of our model θ using this sample. To achieve that task we introduce estimators. An estimator a function of sample.

$$\theta_n = \mu(\{Y_i, X_i\}_{i=1..n})$$

We then set ourselves with an identified model $M = \{F(Y, X; \theta), \theta \in \Theta\}$ w.r.t to $F(Y, X)$. And we consider an estimator θ_n. More over we usually consider the case where the data was generated from that particular model at what we call the true parameter $\theta = \theta_0$. When the data is generated according to that DGP, we will write $\mathbb{E}_0$

1. We say that θ_n is **unbiased** for θ_0 iff $\mathbb{E}_0\theta_n = \theta_0$
2. We say that θ_n is **consistent** for θ_0 iff $\theta_n \xrightarrow{p_0} \theta_0$

We have seen for instance that in the case of the mean, $\theta_n = \frac{1}{n}\sum X_i$ is a consistent estimator. What about consistency and unbiasedness of the variance estimator?

$$\sigma_{2n} = \frac{1}{n}\sum (X_i - \mu_n)^2$$

Beyond point estimates of parameters, we are also interested in forming **confidence intervals** on parameters θ. A $1-\alpha$ confidence interval is a combination of two estimators a_n, c_n (function of the data) such that

$$P(\theta \in [a_n, c_n]) \geq 1 - \alpha$$

where θ is fixed and a_n, c_n are the random variables. See the example for a normally distributed estimator.

We will often consider **asymptotic** confidence interval where we will replace the inequality with a probability limit:

$$P(\theta \in [a_n, c_n]) \xrightarrow{p} 1 - \alpha$$

Finally we will sometimes be interested in conducting hypothesis testing. We will come back to that later.

The case of non iid samples

In many cases we want to model the dependence between variables in our sample. In this case, we need to describe the population in a more complicated way. In general, it might be necessary to provide a DGP for each sample sizes n. This will be the case when modeling spatial correlation or when considering fixed-effects.

3 Ordinary Least Squares

Let's apply our machinery and consider the following model $M = \{Y = X\beta + \epsilon; (\beta, F_{X\epsilon})\}$ where we are interested in β. X is a k dimension random variable and X_n is a $k \times n$ matrix of data.

First question, is the model identified? We need to restrict further our class! Let's use conditional mean independence $\mathbb{E}[\epsilon|X] = 0$. We can then derive identification:

$$\begin{aligned}(\mathbb{E}_0 XX')^{-1}\mathbb{E}_0 XY &= (\mathbb{E}_0 XX')^{-1}\mathbb{E}_0 XX'\beta_0 + (\mathbb{E}_0 XX')^{-1}\mathbb{E}_0 XE_n \\ &= \beta_0 + 0\end{aligned}$$

which shows that β can be recovered from the population data since both $\mathbb{E}_0 XX'$ and $\mathbb{E}_0 XY$ can be computed in the population . We then define the OLS estimator as

$$\beta_n = (X_n'X_n)^{-1}X_n'Y_n$$

which is a function of a given sample (Y_n, X_n).

Is this estimator unbiased?

$$\begin{aligned}\mathbb{E}_0\beta_n &= \mathbb{E}_0 (X_n'X_n)^{-1}X_n'Y_n \\ &= \mathbb{E}_0 (X_n'X_n)^{-1}X_n'(X_n\beta_0 + E_n) \\ &= \mathbb{E}_0 (X_n'X_n)^{-1}X_n'X_n\beta_0 + \mathbb{E}_0 (X_n'X_n)^{-1}X_n'E_n \\ &= \beta\end{aligned}$$

Is this estimator consistent?

$$\begin{aligned}\beta_n &= (X_n'X_n)^{-1}X_n'Y_n \\ &= (X_n'X_n)^{-1}X_n'X_n\beta_0 + (X_n'X_n)^{-1}X_n'E_n \\ &= \beta_0 + (X_n'X_n)^{-1}X_n'E_n \\ &= \beta_0 + \left(\frac{1}{n}X_n'X_n\right)^{-1}\left(\frac{1}{n}X_n'E_n\right)\end{aligned}$$

and so we have that

$$\text{plim}\beta_n = \beta_0 + \left(\text{plim}\frac{1}{n}X_n'X_n\right)^{-1}\left(\text{plim}\frac{1}{n}X_n'E_n\right)$$

then $\text{plim}\frac{1}{n}X_n'E_n = \text{plim}\frac{1}{n}\sum_i x_i\epsilon_i = \mathbb{E}X\epsilon = 0$ (under existence of these limits).

Let's conclude with the asymptotic distribution of the estimator, we look at

$$\begin{aligned}\sqrt{n}(\beta_n - \beta) &= \sqrt{n}(X_n'X_n)^{-1}X_n'E_n \\ &= \left(\frac{1}{n}X_n'X_n\right)^{-1}\left(\frac{1}{\sqrt{n}}X_n'E_n\right)\end{aligned}$$

Table 1: Clustering standard errors

we have that $\frac{1}{n}X_n'X_n \to \mathbb{E}X'X$. Now we should look at the second term. By the central limit theorem we will converge to

$$\frac{1}{\sqrt{n}}X_n'E_n = \frac{1}{\sqrt{n}}\sum_i x_i'\epsilon_i \xrightarrow{d} \mathcal{N}(0, \mathbb{E}X_n'E_nE_n'X_i)$$

if we are in an iid case then $\mathbb{E}\left[E_n'E_n|X_n\right] \to \sigma^2 I_d$ and so

$$\sqrt{n}(\beta_n - \beta) \xrightarrow{d} \mathcal{N}(0, \sigma^2\left(\mathbb{E}XX'\right)^{-1})$$

as we have seen before, we can then use this to construct asymptotic confidence intervals. There are 2 potential sources of complication. First there might be heteroskedasticity, meaning that the variance might be correlated with X. Second there might be correlation in the residuals between observations. This is classic in time series where the error tend to be serially correlated over time. Another case is when errors are serially correlated because of spacial dependence.

The White estimator uses the residuals directly:

$$X'\text{diag}(u_1^2...u_n^2)X$$

3.1 Non-iid samples - clustered standard errors

In this case we want to consider a sampling where observations might be correlated to each other. Going back to our variance expression what we get is

$$\left(\mathbb{E}XX'\right)^{-1}\mathbb{E}XEE'X'\left(\mathbb{E}XX'\right)^{-1}$$

Consider for instance the case of the effect a treatment across villages. We consider the outcome variable

To show the effect of clustering I borrow the example from ...

We see the problems.

4 Parametric Inference

Here we cover two particular methods used in economics: moments based estimation and maximum likelihood.

4.1 Maximum likelihood

Let's consider a parametric model

$$M = \{F(X;\theta); \theta \in \Theta\}$$

we define the maximum likelihood function (which is a statistic) by

$$L_n(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$$

and the log-likelihood function by

$$\ell_n(\theta) = \log L_n(\theta)$$

We finally define the maximum likelihood estimator as

$$\theta_n = \arg\max l_n(\theta)$$

Example: Mean and variance of a normal distribution. Take $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, then $l_n(\theta) = -n\log\sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\bar{X}-\mu)^2}{2\sigma^2}$ where $S^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$ and $\mu = \frac{1}{n}\sum X_i$ and comes from showing that $\sum(X_i - \mu)^2 = nS^2 + n(\bar{X} - \mu)^2$. One can also write the likelihood $l_n(\theta) = -n\log\sigma - \frac{1}{2\sigma^2}\sum_i (Y_i - \mu)^2$ and take FOC.

4 important properties of MLE:

1. MLE is consistent
2. MLE is equivarient (if $\tau = f(\theta)$ then $\tau_n = g(\theta_n)$ is the MLE for τ)
3. MLE is asymptotically normal
4. MLE is efficient (lowest variance estimator)

Overview of consistency

We first introduce the KL distance:

$$D(f, g) = \int f(x) \log(\frac{f(x)}{g(x)}) \mathrm{d}x$$

where it can be shown that $D(f, g) \geq 0$ [1] and $D(f, f) = 0$. Then maximizing $\ell_n(\theta)$ is equivalent to maximizing

$$M_n(\theta) = \frac{1}{n}\sum_i \log\frac{f(X_i;\theta)}{f(X_i;\theta_0)}$$

and by the law of large numbers in converges to

$$\begin{aligned}\mathbb{E}_{\theta_0}\left(\log\frac{f(X_i;\theta)}{f(X_i;\theta_0)}\right) &= \int \log\frac{f(x;\theta)}{f(x;\theta_0)} f(x;\theta_0)\mathrm{d}x \\ &= -D(f(\cdot;\theta), f(\cdot;\theta_0))\end{aligned}$$

[1] By jansen inequality, $D(f, g) = \mathbb{E}_f \log\frac{f(X)}{g(X)} = -\mathbb{E}_f \log\frac{g(X)}{f(X)} \geq -\log\mathbb{E}_f \frac{g(X)}{f(X)} = 0$

and so we see that as $M_n(\theta) \to -D(f(\cdot;\theta), f(\cdot;\theta_0))$, it will be maximized for $\theta = \theta_0$., however showing that the argmax is converging probability requires some conditions. Write $\theta_n = \arg\max_\theta M_n(\theta)$ and

$$\begin{aligned} M(\theta_0) - M(\theta_n) &= M_n(\theta_0) - M(\theta_n) + M(\theta_0) - M_n(\theta_0) \\ &\leq M_n(\theta_n) - M(\theta_n) + M(\theta_0) - M_n(\theta_0) \\ &\leq \sup_{\theta_n} |M_n(\theta_n) - M(\theta_n)| + M(\theta_0) - M_n(\theta_0) \xrightarrow{p} 0 \end{aligned}$$

as long as $\sup |M_n(\theta_n) - M(\theta_n)| \xrightarrow{p} 0$, which requires uniform convergence (ie$M_n(\theta) \xrightarrow{p} M(\theta)$ point-wise is not sufficient). With uniform convergence, as long as θ_0 is the unique maximizer of $M(\cdot)$ we get that $\theta_n \to \theta_0$.

Asymptotic Normality

We first define two new objects. We define the score as

$$s(X;\theta) = \frac{\partial \log f(X;\theta)}{\partial \theta}$$

and the Fisher information as

$$\begin{aligned} I_n(\theta) &= V_\theta \left(\sum_i s(X_i;\theta) \right) \\ &= \sum_i V_\theta \left(s(X_i;\theta) \right) \\ &= n \cdot I(\theta) \end{aligned}$$

The main result is on the asymptotic normality of the MLE estimator:

$$\frac{\theta_n - \theta_0}{\sqrt{1/I_n(\theta_n)}} \xrightarrow{d} \mathcal{N}(0,1)$$

We first show that

$$\mathbb{E}_\theta s(X;\theta) = 0$$

and

$$I = -\mathbb{E}\frac{\partial^2 f(X;\theta)}{\partial \theta^2} = \mathbb{E} s(X_i;\theta) s(X_i;\theta)^t$$

We show the first one:

$$\begin{aligned} \mathbb{E}_\theta s(X;\theta) &= \int \frac{\partial \log f(x;\theta)}{\partial \theta} \cdot f(x;\theta) \mathrm{d}x \\ &= \int \frac{1}{f(x;\theta)} \frac{\partial f}{\partial \theta}(x;\theta) \cdot f(x;\theta) \mathrm{d}x \\ &= \int \frac{\partial f}{\partial \theta}(x;\theta) \mathrm{d}x \\ &= \frac{\partial}{\partial \theta} \int f(x;\theta) \mathrm{d}x = 0 \end{aligned}$$

To get asymptotic normality we then look at

$$0 = \ell_n'(\theta_n) \simeq \ell_n'(\theta) + (\theta_n - \theta)\ell_n''(\theta)$$

and so we get

$$\sqrt{n}\,(\theta_n - \theta_0) \simeq \frac{\frac{1}{\sqrt{n}}\ell_n'(\theta)}{\frac{1}{n}\ell_n''(\theta)}$$

the numerator has mean 0 and variance $I(\theta)$ by the central limit theorem. The denominator by the law of large numbers converges to

$$\mathbb{E}\ell''(\theta) = I(\theta)$$

but we have that

$$\begin{aligned}
\mathbb{E}\ell''(\theta) &= \mathbb{E}\left[\frac{\partial^2 \log f(x;\theta)}{\partial\theta^2}\right] = \mathbb{E}\left[\frac{\partial}{\partial\theta}\frac{1}{f}\frac{\partial f}{\partial\theta}\right] = \mathbb{E}\left[-\frac{1}{f^2}\left(\frac{\partial f}{\partial\theta}\right)^2 + \frac{1}{f}\frac{\partial^2 f}{\partial\theta^2}\right] \\
&= -\mathbb{E}\left[\left(\frac{\partial \log f}{\partial\theta}\right)^2\right] + \mathbb{E}\left[\frac{1}{f}\frac{\partial^2 f}{\partial\theta^2}\right] \\
&= -I(\theta) + \mathbb{E}\left[\frac{1}{f}\frac{\partial^2 f}{\partial\theta^2}\right]
\end{aligned}$$

where we have used that $I(\theta) = \mathbb{E}\left[\left(\frac{\partial \log f}{\partial\theta}\right)^2\right]$ we then show that the second term is 0.

$$\mathbb{E}\left[\frac{1}{f}\frac{\partial^2 f}{\partial\theta^2}\right] = \int \frac{\partial^2 f}{\partial\theta^2}\mathrm{d}x = \frac{\partial^2}{\partial\theta^2}\int f\,\mathrm{d}x = 0$$

so in conclusion we get that $\frac{1}{n}\ell_n''(\theta) \xrightarrow{p} -I(\theta)$ and finally we get that

$$\sqrt{n}\,(\theta_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \frac{1}{I(\theta)})$$

further more we can get an estimator of $I_n(\theta)$ of $nI(\theta)$ and show that

$$\frac{(\theta_n - \theta_0)}{\sqrt{1/I_n(\theta_n)}} \xrightarrow{d} \mathcal{N}(0,1)$$

which allows us to construct confidence intervals.

4.2 Discrete choices and McFadden Multinomial choice

In this section we introduce a very common model of discrete choices. Consider choices according to the following utility:

$$u_i(j) = \beta \log(w) + \gamma X_i + \delta Z_j + \epsilon_{ij}$$

and agent choose favorite alternative. Hence each agent chooses

$$j^*(i) = \arg\max_j u_i(j)$$

under the assumption that ϵ_{ij} is type one extreme value and independent across alternatives we get that

$$Pr[j^*(i) = j] = \frac{\exp\left(\beta \log(w) + \gamma X_i + \delta Z_j + \epsilon_{ij}\right)}{\sum_{j'} \exp\left(\beta \log(w) + \gamma X_i + \delta Z_j + \epsilon_{ij}\right)}$$

Application: Firm with monopsony power

4.3 Dynamic discrete choice

The Bellman principle

There are many situations where decision are dynamic, and made over time. A famous example is the Bus engine paper of John Rust.

Application: Rust bus engine model. http://people.hss.caltech.edu/~mshum/stats/rust.pdf

4.4 Moment based estimation

Imagine that instead of a likelihood, you can construct a function of the sample $Q_n(\theta)$ such that $Q_n(\theta) \overset{u.p.}{\to} Q(\theta)$ and $\theta_0 = \arg\max_{\theta'} Q(\theta')$ (this is an identification condition). It then seems natural to define the following estimator:

$$\theta_n = \arg\max_{\theta} Q_n(\theta)$$

Can we say anything about consistency? Asymptotic normality? we can replicate the argument from the previous section since here we assume that $Q_n(\theta) \overset{u.p.}{\to} Q(\theta)$. Let's then think about the asymptotic normality.

$$\frac{\partial Q_n(\theta_n)}{\partial \theta} = 0 = \frac{\partial Q_n(\theta_0)}{\partial \theta} + \frac{\partial^2 Q_n(\theta_0)}{\partial \theta^2}(\theta_n - \theta_0) + ...$$

so then we can write

$$\sqrt{n}(\theta_n - \theta_0) \simeq -\left(\frac{\partial^2 Q_n(\theta_0)}{\partial \theta^2}\right)^{-1} \sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta}$$

then consider $\frac{\partial^2 Q_n(\theta_0)}{\partial \theta^2} \overset{p}{\to} A_0$ and $\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} \overset{d}{\to} \mathcal{N}(0, B_0)$ we get that

$$\sqrt{n}(\theta_n - \theta_0) \overset{d}{\to} \mathcal{N}(0, A_0^{-1} B_0 A_0^{-1})$$

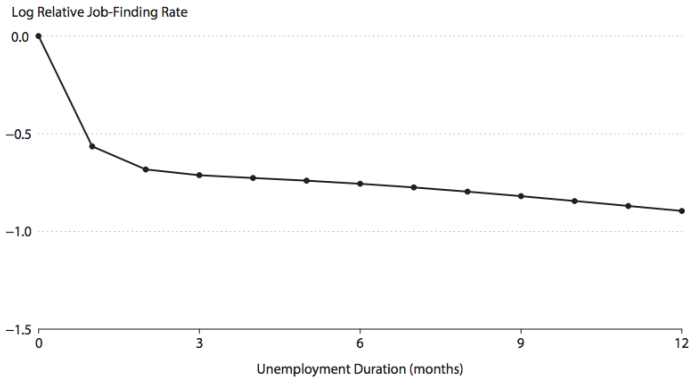
Note that here we do not have the Information matrix equality, hence when compute our C.I. we need to scale the variance using both first and second derivatives.

5 Random effect models

We start by considering the random effect case where:

$$F(X) = \sum_k p_k F_k(X)$$

$$f(x) = \int_\eta f(x|\eta) g(\eta) \mathrm{d}\eta$$



NOTE: The figure displays the negative duration dependence observed in the full sample. The finding rate falls as the duration of unemployment increases.

SOURCE: BLS CPS and authors' calculations.

Figure 1: Unemployment Duration Dependence

which are usually referred to as latent heterogeneity, or finite and infinite mixture models. The first question you might ask is whether in general this is identified? Clearly, in the case where $X \in \mathbb{R}$, we can't separately identify p_k and F_k.

5.1 A running example: unemployment hazard rate

Let's consider as a running example the modeling of unemployment duration. Consider that we are given access to unemployment spells for individuals in a population. This data comes in the form of a list of durations for each individual. Because a given individual might have multiple spells, the data has a panel dimension. Our sample is a list of durations $(\lambda_{i1}, \lambda_{i2}, ...)_{i=1..n}$.

Our first pass would be to consider a simple independent exponential model $Pr[\lambda_{ij} = \lambda] = \gamma \exp(-\gamma\lambda)$. This imposes very strong assumptions on the data: independence, identically distributed and constant hazard rate. When looking at the unemployment data, we find that the constant hazard rate assumption does not seem to hold. From Eubanks and Wiczer, here is a plot of the hazard rate out of unemployment:

Our exponential model would give a constant value for the hazard rate (show this as an example). The first modification is then to allow for a more flexible parametric model, without considering individual heterogeneity per se. For instance, we can consider a finite mixture of exponentials.

$$Pr[\lambda_{ij} = \lambda] = \sum_k p_k \gamma_k \exp(-\gamma_k \lambda)$$

and with enough components this will fit perfectly the cross-sectional distribution of unemployment duration. In this case the probability of multiple duration is given by

$$Pr[\lambda_{i1} = \lambda_1, ..., \lambda_{iJ_i} = \lambda_J] = \prod_j \left(\sum_k p_k \gamma_k \exp(-\gamma_k \lambda_j) \right)$$

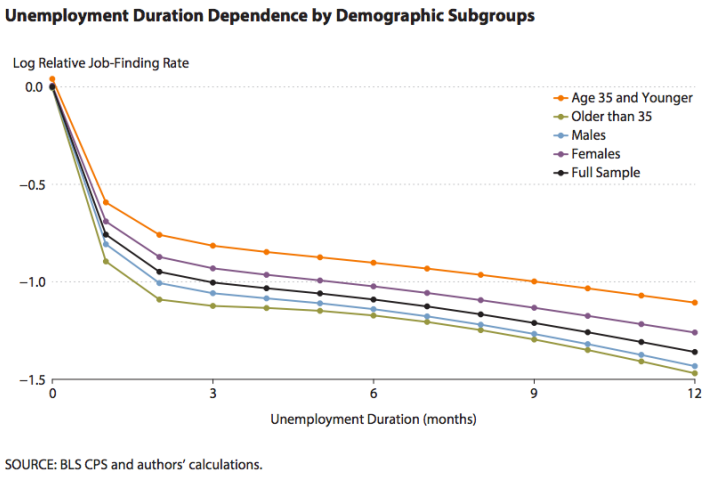


Figure 2: Unemployment Duration Dependence

But this does not tell us about individual heterogeneity, indeed here the mixture over distribution is way to allow for the distribution to be flexible, but nothing tells us whether the change in the hazard rate is due to selection (different people with each their own hazard rate) or due to duration dependence (it gets harder to find a job the longer you are unemployed). The review paper from Eubanks and Wiczer provides a further analysis:

which suggests that we might want to consider a model with heterogeneity at the individual level. A large literature has focused on separating duration dependence from heterogeneity even using single spell data. For the purpose of this lecture we are going to consider the availability of multi-spell data. We then consider the following model:

$$Pr[\lambda_{i1} = \lambda_1, ..., \lambda_{iJ_i} = \lambda_J] = \sum_k p_k \prod_j \gamma_k \exp(-\gamma_k \lambda_j)$$

note the difference between this and the previous model where we had the product around. This model can be extended to include non constant hazard rates using the Weibull distribution instead where you replace $\gamma_k \exp(-\gamma_k \lambda_j)$ by $\rho_k \gamma_k (\gamma_k \lambda_j)^{\rho_k - 1} \exp(-(\gamma_k \lambda_j)^{\rho_k})$:

$$Pr[\lambda_{i1} = \lambda_1, ..., \lambda_{iJ_i} = \lambda_J] = \sum_k p_k \prod_j \rho_k \gamma_k (\gamma_k \lambda_j)^{\rho_k - 1} \exp(-(\gamma_k \lambda_j)^{\rho_k})$$

5.2 Estimation using EM

Consider a finite mixture model such as the one we just described, where γ_k captures the unobserved heterogeneity. And then consider the log likelihood

$$\sum_i \sum_j \log \left(\sum_k p_k \prod_j \gamma_k \exp(-\gamma_k \lambda_{ij}) \right)$$

where our parameters of interests are $p_1...p_k$ and $\gamma_1...\gamma_k$. We could consider directly maximizing this non linear problem with respect to all variables. But this might get difficult. The EM proposes the following two steps:

1. E-step: compute the $q_i(k) = Pr[\alpha_i{=}\alpha_k|y_i, \theta^\tau]$ using the data, the model and the θ^τ guess
2. M-step: choose θ to maximize $\sum_i \sum_k Pr[\eta_i{=}\eta_k|y_i, \theta^\tau] \cdot \log Pr[y_i, \eta_i|\theta]$

let's consider what it means in our example. The E-step is relatively simple if we can easily compute the $Pr[\alpha_i{=}\alpha_k|y_i, \theta^\tau]$ which we can since we know that in our context

$$q_i(k) = Pr[\gamma_i{=}\gamma_k|\lambda_{i1}, \lambda_{i2}, \theta^\tau] = \frac{p_k\gamma_k \exp(-\gamma_k\lambda_{i1}) \exp(-\gamma_k\lambda_{i1})}{\sum_p p_p\gamma_p \exp(-\gamma_p\lambda_{i1}) \exp(-\gamma_p\lambda_{i1})}$$

Then the maximization step consist of maximizing $\sum_i \sum_k Pr[\eta_i{=}\eta_k|y_i, \theta^\tau] \cdot \log Pr[y_i, \eta_i|\theta]$ which here is given by

$$\begin{aligned}\sum_i \sum_k q_i(k) \cdot \log Pr[y_i, \eta_i|\theta] &= \sum_i \sum_k q_i(k) \cdot \left(\log p_k \prod_j \gamma_k \exp(-\gamma_k\lambda_{ij}) - \mu(\sum_k p_k - 1) \right) \\ &= \sum_i \sum_k q_i(k) \cdot \left(\log p_k + J \cdot \log \gamma_k - \sum_j \gamma_k\lambda_{ij} - \mu(\sum_k p_k - 1) \right)\end{aligned}$$

where we need to recover the following parameters $p_1...p_k$ and $\gamma_1...\gamma_k$, taking the FOC we get:

$$\begin{aligned}p_k &= \frac{\sum_i q_i(k)}{\sum_{k'} \sum_i q_i(k')} \\ \gamma_k &= \frac{\sum_k q_i(k) \frac{1}{J} \sum_j \lambda_{ij}}{\sum_{k'} q_i(k')}\end{aligned}$$

Theory behind the EM

We consider a general latent variable model. Denote the augmented data (y_i, α_i) and the log-likelihood model is

$$\begin{aligned}l(y_i) &= \log Pr[y_i|\theta] \\ &= \log \sum_k Pr[y_i, \alpha_i{=}\alpha_k|\theta]\end{aligned}$$

To understand the EM we start with a parameter guess θ^τ and we consider the expression of interest, the log-likelihood. For any k the following is true:

$$\forall k, \qquad \sum_i \log Pr[y_i|\theta] = \sum_i \log Pr[y_i, \eta_i{=}\eta_k|\theta] - \log Pr[\eta_i{=}\eta_k|y_i, \theta]$$

We then take the expectation of the previous expression with weights $Pr[\eta_i{=}\eta_k|y_i, \theta^\tau]$

$$\begin{aligned}\sum_i \log Pr[y_i|\theta] &= \sum_i \sum_k Pr[\eta_i{=}\eta_k|y_i,\theta^\tau]\,(\log Pr[y_i,\eta_i|\theta] - \log Pr[\eta_i|y_i,\theta]) \\ &= \sum_i \sum_k Pr[\eta_i{=}\eta_k|y_i,\theta^\tau]\cdot \log Pr[y_i,\eta_i|\theta] - \sum_i \sum_k Pr[\eta_i{=}\eta_k|y_i,\theta^\tau]\cdot \log Pr[\eta_i|y_i,\theta] \\ &= Q(\theta|\theta^\tau) + H(\theta|\theta^\tau)\end{aligned}$$

The EM algorithm then consists of 2 steps:

1. E-step: compute the $q_i(k) = Pr[\eta_i{=}\eta_k|y_i,\theta^\tau]$ using the data, the model and the θ^τ guess
2. M-step: choose θ to maximize $Q(\theta|\theta^\tau) = \sum_i \sum_k Pr[\eta_i{=}\eta_k|y_i,\theta^\tau]\cdot \log Pr[y_i,\eta_i|\theta]$

The proof that the EM algorithm is always increasing compares the likelihood at θ^τ and $\theta^{\tau+1}$. Expand both with $Pr[\eta_i{=}\eta_k|y_i,\theta^\tau]$ to get the difference equal to:

$$\sum_i \log Pr[y_i|\theta^{\tau+1}] - \sum_i \log Pr[y_i|\theta^\tau] = \big(Q(\theta^{\tau+1}|\theta^\tau) - Q(\theta^\tau|\theta^\tau)\big) + \big(H(\theta^{\tau+1}|\theta^\tau) - H(\theta^\tau|\theta^\tau)\big)\,,$$

where we have that $Q(\theta^{\tau+1}|\theta^\tau) - Q(\theta^\tau|\theta^\tau) \geq 0$ since $\theta^{\tau+1}$ is chosen to maximize that quantity $Q(\theta|\theta^\tau)$. A closer look at $H(\theta|\theta^\tau) - H(\theta^\tau|\theta^\tau)$ reveals that it is minus the Kullback–Leibler divergence between $Pr[\eta_i{=}\eta_k|y_i,\theta^\tau]$ and $Pr[\eta_i{=}\eta_k|y_i,\theta]$:

$$\begin{aligned}H(\theta^{\tau+1}|\theta^\tau) - H(\theta^\tau|\theta^\tau) &= \sum_i \sum_k Pr[\eta_i{=}\eta_k|y_i,\theta^\tau]\cdot \log \frac{Pr[\eta_i|y_i,\theta^\tau]}{Pr[\eta_i|y_i,\theta^{\tau+1}]} \\ &= D_{KL}(Pr[\eta_i|y_i,\theta^\tau], Pr[\eta_i|y_i,\theta^{\tau+1}]) \geq 0\end{aligned}$$

and hence will always be negative. This shows that the likelihood increases at each step.

5.3 On the identification of Hazard rate models

For single and multi-spell data. TBD

5.4 Random effect versus fixed effect

TBD (assumptions on the distribution...).

5.5 Dynamic discrete choice with unobserved types

An important application of random effect in economics has been in the estimation of dynamic discrete choice models. Such models are concerned with modeling dynamic decisions of individuals, see Aguirregabiria and Mira (2010) for a survey of these methods.

In each period agents can choose a discrete action $a_t \in [1..J]$ and we call s_{it} the state space of the agent. Agents preference are given by

$$\mathbb{E}_t \sum_{\tau=t}^{T} \beta^{\tau-t} u(a_{i,\tau}, s_{i,\tau})$$

and the Bellman principal gives us that the optimal decision rule is of the form:

$$V(s_{it}) = \max_{a \in A} \left\{ U(a, s_{it}) + \beta \int V(s_{i,t+1}) dF(s_{i,t+1}|a, s_{it}) \right\}$$

and where the choice specific value is often called the Q-value and is given by

$$Q(a, s_{it}) = u(a, s_{it}) + \beta \int V(s_{i,t+1}) dF(s_{i,t+1}|a, s_{it})$$

The state space is often composed of observed and unobserved variables $s_{it} = (x_{it}, \epsilon_{it})$ and usually we observe a pay-off function $y_{it} = Y(a_{it}, x_{it}, \epsilon_{it})$. The model then specifies the $u(\cdot)$ function. The data is then a sequence $\{a_{it}, x_{it}, y_{it} : i = 1..N; t = 1..T_i\}$.

we can then write the likelihood of the model

$$\begin{aligned} l_i(\theta) &= \log Pr[a_{it}, y_{it}, x_{it} : t = 1..T_i|\theta] \\ &= \log Pr[a_{i1}, y_{i1}, x_{i1}|\theta] \prod_{t=2}^{T_i} Pr[a_{it}, y_{it}, x_{it}|a_{it-1}, y_{it} - 1, x_{it-1}\theta] \\ &= \log Pr[a_{i1}, y_{i1}, x_{i1}|\theta] + \sum_{t=2}^{T_i} \log Pr[a_{it}, y_{it}, x_{it}|a_{it-1}, y_{it} - 1, x_{it-1}\theta] \end{aligned}$$

Real example. Kean and Wolpin (1994) consider a model of labor choices using this Framework. Starting at age 16, agents choose among 5 alternatives, staying at home $a_{it} = 0$),going to school ($a_{it} = 4$) or one of 3 occupations white collar ($a_{it} = 1$), blue collar ($a_{it} = 2$), military ($a_{it} = 3$). Utility functions have the form $U(a, s_{it}) = \omega_i(a) + W_{it}(a)$, where the wage W_{it} is 0 for $a_{it} = 0, 4$.

This is of course without any permanent unobserved heterogeneity. But we have seen how to augment a likelihood framework to include random unobserved discrete heterogeneity. We then assume that there are K unobserved types which drive potentially the law of motion $F(s_t|a, s_{t-1}, k)$ and of course the pay-off function Y and the preference u.

$$\begin{aligned} l_i(\theta) &= \sum_k p_k l_i(\theta, k) \\ &= \sum_k p_k \log Pr[a_{it}, y_{it}, x_{it} : t = 1..T_i|\theta, k] \end{aligned}$$

which can be estimated by direct minimization, or using the EM.

6 Linear regression with many regressors

This topic relates to model selection, or regularization of the linear regression. This is relevant in particular when the number of regressor is very large when compared to the number of observations. Many datasets come with a large number of observables that can be used to explain a variable of interest. So we consider the following linear model

$$Y = X'\beta + \epsilon$$

where we have p regressor and we consider a sample $p \ll n$. It is clear that as soon as $p > n$, trying to directly solve the system of equation will lead to multiple possible solution. As you can see this is not the same concept as identification since if p is actually fixed in the population, then the model will be formally identified, however in a sample, the OLS estimator might give extremely poor estimates, estimates with a very high variance (that will depend a lot on the picked sample).

6.1 Stepwise selection

Go through the regressors and add them one by one, choose the best using R^2 then select using cross-validation.

6.2 Ridge Regression

One approach is to penalize the coefficients of the regression. This is referred to as a Ridge regression or shrinkage. The estimator is defined as:

$$\sum_i \left(y_i - \beta_0 - \sum_j \beta_j x_{ij} \right) + \lambda \sum_j \beta_j^2$$

This is not obvious why this would perform any better. We know in particular that the OLS with $\lambda = 0$ is the best linear unbiased estimator. However we might want to trade a bit of bias to lower the variance.

6.3 Lasso

The lasso proposes a similar approach but instead penalizes the absolute deviation of the parameter instead of the square of the parameter.

$$\sum_i \left(y_i - \beta_0 - \sum_j \beta_j x_{ij} \right) + \lambda \sum_j |\beta_j|$$

We can compare Lasso and Ridge in the simple case where $p = n$ and X is a diagonal matrix. In this case it can be shown that OLS gives

Application: Elena Manresa's paper on R&D effects

6.4 Principal component analysis

The idea of using principal component analysis is to recover underlying explanatory factors in the regressors before even starting the regression analysis. Imagine we have a prior that the X can be represented as a lower dimensional linear combination of variables. Then we want to look for a vector α such that the following variance is maximized:

$$X'\alpha$$

To find such value with conduct an SVD decomposition of X.

$$\alpha = \arg\max \frac{w'X'Xw}{w'w}$$

We want to maximize

$$\frac{1}{2}\alpha'\Sigma\alpha - \lambda(\alpha'\alpha - 1)$$

where FOC give us

$$\alpha'\Sigma = \lambda\alpha$$

not that then the objective is λ so we should pick the largest eigen value! We do this in a first step, perhaps select the k largest eigen values. We then use this in a regression

caveats: the dimension reduction is not driven by the relationship to Y. This can be potentially problematic, especially if the residual is very large.

7 Non parametric regression

In this section we are going to look at non-parametric model:

$$M = \{Y = f(X) + \epsilon; f \in \mathbb{R}^{\mathbb{R}}, \mathbb{E}[\epsilon|X] = 0\}.$$

We can see right away that the difficulty here will be that the parameter of interest is infinite dimensional. The first question we ask is wether the model is identified. We see however that this comes out naturally since

$$\mathbb{E}[Y|X = x] = \mathbb{E}[f(X) + \epsilon|X = x] = f(x)$$

we then want to consider potential estimators for this model. There are three classes of estimators we can consider:

- A kernel estimator uses a local approach and fits locally weighted regressions:

$$f_n(x) = \frac{\sum_i Y_i \cdot K_h(X_i - x)}{\sum_i K_h(X_i - x)}$$

- A sieve estimator will approximate a function using a finite number of components:

$$f_n(x) = \sum_k w_{nk} g_k(x) \text{ where } w_{nk} = \arg\min \sum_i ||Y_i - \sum_k w_{nk} g_k(X_i)||$$

- A tree method tries to approximate the function with piece wise constant functions

$$f_n(x) = \sum_k c_k I[x \in R_m]$$

7.1 Choice of tuning parameter

In each of the three methods the econometrician needs to choose a smoothing parameters. a bandwidth h for the kernel estimator, the number of components K for the sieve or for the regression tree. Given a sample (Y_n, X_n), we see that we can always choose such tuning parameter to fit the data as well as we want. Indeed consider adding more and more $g_k(x)$ to your analysis and when $K = N$ you will get a just invertible matrix.

The parameter can then generate over-fitting in sample. This can be thought of in term of a bias-variance tradeoff for the estimator. We compute the mean square error of a given estimator:

$$MSE(\theta_n) = \mathbb{E}(\theta_n - \theta)^2 = \underbrace{Var(\theta_n)}_{\text{variance}} + \underbrace{(\mathbb{E}(\theta_n - \theta))^2}_{\text{Bias}}$$

This is hard to get in general since θ is not known. One way to resolve this problem is to use cross-validation.

7.2 Kernel estimator

We first look at the Kernel estimator:

$$f_n(x) = \frac{\sum_i Y_i \cdot K(\frac{X_i - x}{h})}{\sum_i K(\frac{X_i - x}{h})}$$

One can derive the asymptotic properties of this estimator. This might be going beyond the scope of this notes. Let's report the result:

$$\sqrt{nh}\left(g_n(x) - g(x) - h^2\kappa_2 B(x)\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{R(k)\sigma^2(x)}{f(x)}\right).$$

We notice that as $h \to 0$, $nh \to \infty$, we get consistency of the estimator. Two points are relatively interesting. The first is that the estimator has a bias for any positive value of h.

$$B(x) = \frac{1}{2}g''(x) + \frac{g'(x)f'(x)}{f(x)}$$

this seems intuitive, when h is to large, it is impossible for the estimator to pick large variations locally in the $g()$ function. Note that we get he following MSE:

$$h^4\kappa_2 B(x) + \frac{R(k)\sigma^2(x)}{nhf(x)}$$

and so minimizing the MSE gives an optimal rate for $h \sim n^{-1/5}$, plugging this back in gives an optimal convergence for the estimator of $n^{2/5}$ versus $n^{1/2}$for parametric estimators.

7.3 Sieve Estimator

suppose that you have a base of function such that $m(x) - \beta' p^K(x) \to 0$ for all x. Then we will have convergence in $O_p(K/n + K^{-2\alpha})$. Note that again here you will get for any fixed K a bias. In the case of polynomial approximation to a smooth function in C^∞, then we get that $\alpha = 1$.

7.4 Regression tree

Remember the structure of the repression tree:

$$f_n(x) = \sum_k c_k I[x \in R_m]$$

The tree is constructed using binary splitting. Start with the entire support and choose a splitting rule (for instance pick a given regressor X_j and find a threshold s such that the MSE is minimized). Then repeat this procedure until each group has a number of observations less than some number $\underline{k}$.

One problem is that some early splits can deliver small RSS reductions but lead to large gains letter on. Hence it can be very costly to stop with high RSS rule when doing binary splitting. Example of square with blocks.

Common strategy is then to fit a very "deep" tree, then go back and "prune" it. We then compute for each sub-tree $T \subset T_0$:

$$\sum_{m=1}^{T} \sum_{x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

and for a given α there is an optimal sub-tree. α can then be picked using cross-validation (k-fold or leave-one-out).

7.5 Semi-parametric estimator

Let's look at index models.

$$Y = g(X'\beta) + \epsilon$$

in this case one can achieve $\sqrt{n}$ consistency for the parameter β.

Mansky maximum score estimator

Consider the model

$$Y_i = 1[X_i'\beta + u_i > 0]$$

when the distribution of u_i is a normal distribution then we have a the probit model. When it is type 1 extreme value we have the logit model. But what if you are not willing to make a distributional assumption? Mansky makes the following observation on average, $Y_i < Y_j$ should be associated with $X_i'\beta < X_j'\beta$. He then proposes the following estimator:

$$\beta_n = \arg\max \sum_{i>j}$$

8 Boostrap

Compute confidence intervals and critical values can be tedious when using asymptotic formulation. If we could draw directly from the population we could conduct a Monte-Carlo exercise and recover the distribution of the estimator. In this section we consider such an approach by sampling from the available sample. Considering a given sample

$Y_1..Y_n$, there are two main re-sampling approach. The first is to re-sample n elements from $(Y_1..Y_n)$ with replacement, the second is to sample $m < n$ from $(Y_1..Y_n)$ without replacement. In both approaches the goal is generate draws from a distribution that reassembles as much as possible to the population distribution.

The theory behind the bootstrap The data is assumed to be independent draws from $F_0(x) = F_0(x, \theta_0)$ and we consider a statistic $T_n = T_n(X_1..X_n)$. The distribution of T_n is denoted by $G_n = G_n(t, F_0) = P_0[T_n \leq t]$. Asymptotic theory relies on G_∞, instead the bootstrap relies on plugging in an estimate of F_0 and uses $G_n(\cdot, F_n)$. Taking B samples with replacement from F_n, computing $T_{n,b}$ in each, we can construct

$$\hat{G}_n(t, F_n) = \frac{1}{B} \sum_b \mathbf{1}[T_{n,b} \leq t]$$

then what we need for the bootstrap procedure to be asymptotically valid is that

$$G_n(t, F_n) \xrightarrow{p} G_n(t, F_0)$$

uniformly in t. This requires *smoothness* in F_0 as well as in $G_n(\cdot, \cdot)$ and consistency of F_n for F_0. In general we get that if we have $\sqrt{n}$ asymptotic convergence to G_∞, then both $G_n(t, F_0)$ and $G_n(t, F_n)$ do so and so they are also close to each other:

$$G_n(t, F_0) = G_n(t, F_n) + O(N^{-1/2})$$

which provides no gain when compared to asymptotic standard error besides the simplicity of the computation.

Parametric Bootstrap Note that the goal is to approximate F_0 and hence F_n is a good candidate however one can use $F(\cdot, \theta_n)$ where θ_n is a consistent estimator of θ_0. This is referred to as the parametric bootstrap.

Asymptotic refinement It can be shown that in the case where T_n is asymptotically pivotal, meaning that is does not depend on the parameters, then the bootstrap achieves:

$$G_n(t, F_0) = G_n(t, F_n) + O(N^{-1})$$

The idea here is that one can get a better approximation of the finite sample distribution. At every N, $G_n(t, F_n)$ is closer to $G_n(t, F_0)$ than $G_\infty(t, F_0)$. This can be shown using the Edgeworth expansion which expands $G_n(z)$ as a function of $n^{-\frac{1}{2}}$.

$$\begin{aligned} G_n(t, F_n) - G_n(t, F_0) = &[G_\infty(t, F_n) - G_\infty(t, F_0)] \\ &+ \frac{1}{\sqrt{n}} [g_1(t, F_n) - g_1(t, F_0)] + O(n^{-1}) \end{aligned}$$

and then $G_\infty(t, F_n) - G_\infty(t, F_0) = 0$ if T_n is asymptotically pivotal, and $g_1(t, F_n) - g_1(t, F_0) = O(n^{-1/2})$ delivering an overall $O(n^{-1})$.

Failure of bootstrap: One example of the failure even when the estimator is asymptotic normal is the nearest neighbor estimator (Abadie and Imbens 2008). It is shown that the variance of the bootstrap is either too small or too large. Another example is the estimation of the median.

Bias correction using bootstrap The bootstrap can be used to correct for the bias of an estimator. In many applications the exact form of the bias $\mathbb{E}_0(\theta_n - \theta_0)$ is not known, however if we consider $\bar{\theta}_n^*$, the expectation across bootstraps replications, then it gives us an estimate of the bias. We can then consider a bias-corrected estimate $\theta_n^{BR} = \theta_n - (\bar{\theta}_n^* - \theta_n)$.

Non iid samples There will cases where the data is not exactly iid. For instance there might be weak spatial correlation. In this case, one might want to bootstrap by resampling clusters of data to replicate the dependence. More on this later.

9 Fixed effect model

Most of the data-generating processes we have considered until now were cross-sectional, in the sense that the population was some distribution $F_x(x)$ and a sample was n draws from this distribution. However, many datasets provide repeated measures for each individual. For instance longitudinal survey track individuals over multiple periods. The data generating provides us with a sequence of variables $(Y_1, X_1 ... Y_T, X_T)$ for $t = 1..T$. Some approaches focus on short T and other on asymptotic as $T \to \infty$.

Given that the data provides multiple measurements, we might want to start thinking about individual specific heterogeneity. They are different ways of addressing this problem. In particular people refer to random effect and fixed effect approaches. In all cases we are going to consider modeling unbosbserved heterogeneity with individual specific unobserved characteristics. Consider a DGP $F(Y|X)$, we will then let this DGP be different for different i inddividuals and denote it in the following way $F_i(Y|X)$ or $F(Y|X, \alpha_i)$. Two main approaches are considered.The first one is to treat α_i as a random variable and model the distribution it is coming from such asa $F(Y|X) = \int F(Y|X, \alpha_i) g(\alpha_i) \mathrm{d}\alpha_i$. This is referred to as **random effect**. The second approach considers the α_i as parameters of the model that have to be estimated, these are called **fixed-effect** approaches. Note that in the fixed effect approach the number of parameters α_i grows with the sample size, whereas it might not be the case for random effect.

We start by treating the fixed-effect model in the general non linear framework. In this case we consider individual heterogeneity as parameters. Our data is a vector $(Y_1 ... Y_T)$ for $i = 1..n, t = 1..T$. We then have a likelihood model $\ell_i(y_i) = \ell(y_i; \theta, \alpha_i)$ and our parameter set grows with the sample size. Our MLE is given by:

$$(\hat{\theta}, \hat{\alpha}_i) = \arg\max \frac{1}{n} \sum_i \sum_t \ell(y_{it}; \theta, \alpha_i).$$

The gains is that know we do not have to model the relationship between α_i and the rest of the model. What this means is that we can't apply our inference framework based on finite parameter inference. In finite T framework, in general the difficulty in

estimating α_i for each individual will contaminate the estimation of θ, making it biased even as $n \to \infty$.

Incidental parameter bias example

We consider a simple example adapted from Chamberlain. Consider the following model:

$$y_{it} \sim \mathcal{N}(\alpha_i, \sigma^2),$$

for which we can write our MLE estimator

$$\ell_i = -\frac{1}{2}\log\sigma - \frac{1}{2}\frac{(y_{it} - \alpha_i)}{\sigma^2}$$

which gives $\alpha_i = \frac{1}{T}\sum y_{it}$ and $\hat{\sigma} = \frac{1}{nT}\sum\sum(y_{it} - \bar{y}_i)^2$ and then

$$\begin{aligned}
\mathbb{E}\hat{\sigma} &= \mathbb{E}(y_{it} - \frac{1}{T}\sum_{t'} y_{it'}) \\
&= (1 - \frac{1}{T})^2\sigma^2 + \frac{1}{T}\sigma^2 \\
&= \sigma^2 - \frac{\sigma^2}{T}
\end{aligned}$$

we notice that in this case, we do not recover σ^2 exactly at fix T even if $N \to \infty$. This is referred to has the incidental parameter bias. In general, for large enough T and under smoothness conditions, it will be the case that θ_n will be centered at θ_T, $\sqrt{nT}\,(\theta_n - \theta_T) \xrightarrow{d} \mathcal{N}(0, \Omega)$ where

$$\begin{aligned}
\theta_T &\equiv \arg\max \lim_{n\to\infty} \frac{1}{n}\sum_i \mathbb{E}\sum_t \log f\left(y_{it}|\theta, \hat{\alpha}_i(\theta)\right) \\
\hat{\alpha}_i(\theta) &\equiv \arg\max_\alpha \sum_t \log f(y_{it}|\theta, \alpha)
\end{aligned}$$

and we get that:

$$\theta_T = \theta_0 + \frac{B}{T} + O(\frac{1}{T^2})$$

Note that this is quite sever since even in the case where $n/T \to \rho$ we get that

$$\sqrt{nT}(\theta_n - \theta_0) = \sqrt{nT}(\theta_n - \theta_T) + \sqrt{nT}\frac{B}{T} + O\left(\sqrt{\frac{n}{T^3}}\right) \xrightarrow{d} \mathcal{N}(B\sqrt{\rho}, \Omega)$$

9.1 Marginal effect

It is often of interest to consider marginal effects, such as some averaging of the parameters and the fixed effects

$$\sum_i m(y_{i,}\alpha_i, \theta)$$

9.2 Group-fixed effect estimators

A last approach is to treat the heterogeneity as discrete in estimation (Bonhomme, Lamadon, and Manresa, 2016). We still consider a fix-effect world with $f(y_i|\theta, \alpha_i)$ which we are interested in, however, in estimation we are going to classify individuals into groups, and then treat these groups as observed types in a second step. For this to work we then assume that we can find individual moments h_i such that $h_i = \varphi(\alpha_i)+$

10 Linear Panel

The simple linear model can be written as

$$y_{it} = \beta X_{it} + \alpha_i + \epsilon_{it}$$

sufficient conditions for identification are that $\mathbb{E}(\epsilon_{it}|X_{it}, \alpha_i) = 0$, but actually even weaker conditions such as $\mathbb{E}(\epsilon_i|X_i) = 0$ are ok, as we can see that we can construct a first difference estimator.

A common estimator is the Within Group estimator:

$$\Delta y_{it} = \beta \Delta X_{it} + \Delta \epsilon_{it}$$
$$DY_i = \beta DX_i + DE_i$$

where

$$D = \begin{bmatrix} -1 & 1 & & 0 \\ & \ddots & \ddots & \\ 0 & & -1 & 1 \end{bmatrix}$$

running the regression in differences or allowing for dummies for each individual is mathematically identical.

$$\beta_n = \left(\sum_i (DX_i)' X_i D \right)^{-1} \sum_i DX_i Dy_i$$

10.1 Omitted variable

one case where this can be very useful is in the case of the presence of an omitted variable.

$$y_{it} = \beta X_{it} + \gamma Z_i + \epsilon_{it}$$

you can think for instance of the case where Z_i is a permanent ability, which is unobserved, but of course correlated with the other regressors X_{it}. In this case we can see that the OLS estimator of y_{it} on x_{it} will give:

$$\beta^{OLS} = \frac{cov(y_{it}, x_{it})}{var(x_{it})} = \frac{cov(\beta X_{it} + \gamma Z_i, x_{it})}{var(x_{it})} = \beta + \gamma \frac{\sigma_{xz}}{\sigma_x^2}$$

in the presence of repeated measure, we can use a first difference estimator using the fact that

$$\mathbb{E}[x_{it} (\Delta y_{it} - \beta \Delta x_{it})] = 0$$

10.2 Measurement error in regressors

here we consider the following model

$$y_{it} = \beta x_{it}^{\dagger} + \epsilon_{it}$$
$$x_{it} = x_{it}^{\dagger} + u_{it}$$

This can be estimated using the following moments:

$$\mathbb{E}\left[x_{is}\left(y_{it} - \beta x_{it}\right)\right] = 0$$

in the case here we also have a correlated fixed effect then we use the following moments:

$$\mathbb{E}\left[x_{is}\left(\Delta y_{it} - \beta \Delta x_{it}\right)\right] = 0$$

10.3 Autoregressive model

Consider the following model:

$$y_{it} = \rho y_{it-1} + \alpha_i + \epsilon_{it}$$

with $\mathbb{E}[\epsilon_{it}|y_i^{t-1}, \alpha_i] = 0$. We can start by looking at the OLS estimator:

$$\begin{aligned}
\rho^{OLS} &= \frac{cov(y_{it}, y_{it-1})}{var(y_{it-1})} \\
&= \frac{cov(\rho y_{it-1} + \alpha_i + \epsilon_{it}, y_{it-1})}{var(y_{it-1})} \\
&= \rho + \frac{1}{1-\rho} Var(\alpha_i)
\end{aligned}$$

next we can look at the first different estimator

$$\begin{aligned}
&= \frac{cov(\Delta y_{it}, \Delta y_{it-1})}{var(\Delta y_{it-1})} \\
&= \frac{cov(\rho \Delta y_{it-1} + \Delta \epsilon_{it}, \Delta y_{it-1})}{var(\Delta y_{it-1})} \\
&= \rho + \frac{cov(\Delta \epsilon_{it}, \Delta y_{it-1})}{var(\Delta y_{it-1})} \\
&= \rho - \frac{\sigma_{\epsilon}^2}{var(\Delta y_{it-1})}
\end{aligned}$$

Using instruments:

$$\mathbb{E} y_i^{t-2}\left(\Delta y_{it} - \rho \Delta y_{it-1}\right)$$

which is consistent even at fixed T.

Table 1
Firm Money Demand Estimates
Sample period 1986–1996

	OLS Levels	OLS Orthogonal deviations	OLS 1st-diff.	GMM 1st-diff.	GMM 1st-diff. m. error	GMM Levels m. error
Log sales	.72	.56	.45	.49	.99	.75
	(30.)	(16.)	(12.)	(16.)	(7.5)	(35.)
Log sales	−.02	−.03	−.03	−.03	−.03	−.03
×trend	(3.2)	(9.7)	(4.9)	(5.3)	(5.0)	(4.0)
Log sales	.001	.002	.001	.001	.001	.001
$\times$trend^2	(1.2)	(6.6)	(1.9)	(2.0)	(2.3)	(1.4)
Sargan				.12	.39	.00
(p-value)						

All estimates include year dummies, and those in levels also include industry dummies. t-ratios in brackets robust to heteroskedasticity & serial correlation. N=5649. Source: Bover and Watson (2005).

Figure 3: Firm money demand

10.4 Example, Firm money demand

11 Recap of important concepts:

1. Definition of a probability space
2. Notion of convergence (probability, distribution)
3. Central limit theorem
4. Definition of a model and identification
5. Concept of samples, estimation/inference, CI
6. Properties of estimators, consistency, unbiasedness, $\sqrt{n}$ convergence
7. NP regression, tuning parameter, bias-variance trade off, cross-validation
8. Unobserved heterogeneity with repeated measurement
   (a) the incidental parameter problem
   (b) Random effect versus Fixed effect

References

Aguirregabiria, V., and P. Mira (2010): "Dynamic discrete choice structural models: A survey," *J. Econom.*, 156(1), 38–67.

Bonhomme, S., T. Lamadon, and E. Manresa (2016): "Discretizing Unobserved Heterogeneity," Discussion Paper 1536.

A Additional notes

A.1 Integration

Through the document we use a notion of integration with respect to a measure P. Let's define here precisely what is meant by that.

As in most textbooks we start by defining the integral for simple functions. Simple functions are function that take a finite number of values. For instance consider a partition of E_j of Ω with $E_j \in \mathcal{F}$ then the following function is simple

$$\phi(x) = \sum_{j=1}^{n} \mathbf{1}[x \in E_j] c_j$$

and then we define the integral for such function as

$$\int \phi \mathrm{d}P = \sum_j c_j \cdot P(E_j)$$

and then for any positive function we define the integral as

$$\int f \mathrm{d}P = \sup \left\{ \int \phi \mathrm{d}P \text{ s.t. } \phi \text{ simple and } \phi \leq \mathrm{f} \right\}$$

finally any function f can be written as the difference of two positive function $f = f^+ - f^-$ and so the integral definition follows from there.

A.2 o and o_p notations

$X_n = o(R_n)$ means that $X_n = Y_n R_n$ and $Y_n \xrightarrow{p} 0$

also we have that $(1+x)^\alpha = 1 + \alpha x + o(x^2)$

A.3 KL divergence

Let's show that it is always positive

$$\begin{aligned} KL(p,q) &= \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x \\ &= \mathbb{E}_p \log \frac{p(x)}{q(x)} \\ &= -\mathbb{E}_p \log \frac{q(x)}{p(x)} \\ &\geq -\log \mathbb{E}_p \frac{q(x)}{p(x)} \\ &= -\log \int q(x) \mathrm{d}x = 0 \end{aligned}$$

A.4 Linear algebra refresher

- Matrices, product, rank, eigen value decompositions

A.5 Bellman principle of optimality

TBD