CORVINUS UNIVERSITY OF BUDAPEST

Department of Mathematical Economics and Economic Analysis

STATISTICS, ECONOMETRICS, DATA ANALYSIS

Lecture Notes

János Vincze

ISBN: 978-963-503-815-2

Budapest, December 2019

Contents

1 Introduction

These lecture notes are for Economics PhD students at the Corvinus University of Budapest, but can be used equally by any graduate student interested in modern econometrics and its relationship to general statistics. It is divided into five main sections. The first introduces some general concepts of theoretical statistics, including Bayesian ideas. Many of these ideas appear in econometrics textbooks, but some of them is ominously missing. Basic (philosophical) questions of statistics are usually not treated in those, though some appreciation of them should be useful for any practicing econometrician. The next two sections cover material that can be found in most (non-time series) econometrc textbooks. Here I stress the difference between two apparoaches: the data description style of classical regression analysis and the causal estimation centered econometric approach. The following section introduces statistical learning, an area little known for most economists at the present. My conviction is that its knowledge will be more and more crucial in the future. The final section is assigned to time series analysis, mostly dealing with traditional time domain approaches, but making an unusual, for econometric texts, foray into the frequency domain and wavelet methods. Again, I believe that the latter will be important in the future, and the former is a stepping stone to the latter.

The bookis a textbook and as such a compendium. It does not contain new material, at most a few examples or cases. It is based mostly on other textbooks, but selected from a rather wide range. At the end of each section those texts I used most extensively are listed, in all areas these should be consulted if someone wants to have a deeper understanding of the issues involved. The present text aims at a wide coverage, rather than an in-depth one.

2 Classical statistics

2.1 The classical statistical problem

We have n observations on p variables. We assume that there exists a family of distributions parameterized by θ,

$$F(X \mid \theta)$$

where $\theta \in \Theta$, and there exists a specific "true" $\widehat{\theta}$. Then x (the observed sample) is a realization from this distribution. Thus X is a matrix of random variables, and we have available a specific realization x (the sample), which we observe.

The goal is to derive statements about the true $\widehat{\theta} \in \Theta$.

For an example suppose we measure the height of 2000 Hungarian citizens. The assumption is that the selection of these people was random, and there is a random variable with cumulative distribution function (cdf) F that can be called "the height of a living Hungarian". The simplest and most frequent assumption is that $X_1....X_n$ have the same ($F_i = F$) distribution, and they are independent pairwise:

$$F(x_1, x_2, ...x_n) = F_1(x_1)F_2(x_2)...F_n(x_n) = F(x_1)F(x_1)...F(x_n).$$

This sample is called i.i.d. (independently identically distributed). So far we haven't introduced enough restrictions to enable us to speak about "parameters". But, if, in addition, we assume that F is normal, then $E_F(X) = \mu$,and $var_F(X) = \sigma^2 > 0$ completely determine the population distribution, and we have a parametric problem. Classical statistics mostly, but not exlusively, was concerned with parametric problems. In general we can define a statistic T as any measurable function of X, thus $T(X)$ is also a random variable (possibly multivalued), with realization t.

Another possible interpretation of population Another possible interpretation is that the population consists of the almost 10 millions of Hungarians living today. If sampling were conducted with replacement then the sample is also i.i.d. In that case the parameter of interest is simply the average height, and we can infer it, in principle, by observing everyone. The problem is statistical only because it is too costly to observe everyone.

A frequent use of statistics is to forecast the outcome of an election based on exit poll data. Here the goal is to give an estimate of the actual vote of a finite set of individuals, not the potential vote of an infinite potential population. It is similar to the problem of taking a sample of, say a bunch, of bullets, and determine what percentage of the bunch is faulty. Because in this case sampling implies destroying the bullets increasing the sample size would be counterproductive, though it would lead to the truth eventually.

Most econometric investigations assume an infinite population, and strive for more than just establishing some contingent facts about the present or the past. In each particular case one has to decide which interpretation is sensible.

2.2 First approach to telling something about the parameters: point estimation

Let us try to find statistics that "determine" the unknown μ and σ^2 in the above example! In general determining exactly the true parameters is not possible. Thus this is not a well-defined mathematical problem. Classical statisticians' informal purpose is "to get as close as possible" to the true parameters in some sense. There exist basically three ways to achieve this goal: point estimation, interval estimation, and testing.

Point estimation aims at giving the best possible "single" estimate for an unknown parameter (an element of Θ). There is no unequivocal meaning to the expression: "some *statistic is an estimate of parameter* θ". However, informally this expression is used customarily.

2.2.1 Formal properties of estimators

Definition: A statistic (an estimator) is called an unbiased estimator of θ , if

$$E_\theta T(X) = \theta.$$

Unbiasedness seems to be reasonable: on average the estimator returns the true parameter.

Proposition: $E(\overline{X_n}) = \mu$ (the sample average is an unbiased estimate of the population mean), where

$$\overline{X_n} = \left(\frac{1}{n} \sum_{i=1}^{n} X_i \right).$$

There are many unbiased estimates. For instance if $\sum \lambda_i = 1$

$$\overline{X}_{n\lambda} = \sum \lambda_i X_i$$

is unbiased. But what about their variance? It is

$$var(\overline{X}_{n\lambda}) = E(\overline{X}_{n\lambda} - \mu)^2.$$

Definition: An unbiased estimator is efficient if it has minimum variance among unbiased estimators.

Proposition: The sample mean is efficient for the population mean, and its variance is $\frac{\sigma^2}{n}$. It is plausible to estimate the population variance with its sample counterpart:

$$s_u^2 = \sum_i \frac{1}{n}(X_i - \overline{X_n})^2.$$

However, it turns out that it has a little fault.

Proposition: $E(s_u^2) = \frac{n-1}{n}\sigma^2$,in other words, this estimator is biased downwards. The bias can be easily corrected:

$$s^2 = \frac{n}{n-1}s_{en}^2 = \frac{1}{n-1}\sum(X_i - \overline{X}_n)^2.$$

$$E(s^2) = \sigma^2.$$

A strange example: Poisson distribution A Poisson distribution describes the number of occurrences of a random phenomenin per unit of time. Its probability mass function is

$$p_n = \exp(-\lambda)\frac{\lambda^n}{n!}.$$

Suppose we observe the phenomenon n times during a unit interval. We look for the unbiased estimator of λ, as a statistic $T(n)$. Unbiasedness requires that

$$\begin{aligned}\sum_{n=0}^{\infty}\exp(-\lambda)\frac{\lambda^n}{n!}T(n) &= \lambda,\\ \sum_{n=0}^{\infty}\frac{\lambda^n}{n!}T(n) &= \lambda\exp(\lambda).\end{aligned}$$

The left-hand side must be the Taylor-series expansion of the right hand side around $\lambda = 0$. Therefore

$$\begin{aligned}T(n) &= \frac{\partial^n(\exp(\lambda)\lambda)}{\partial n^n}\\ T(n) &= n\end{aligned}$$

One can derive the unbiased estimator of λ^2 with the same method:

$$\begin{aligned}
\sum_{n=0}^{\infty} \exp(-\lambda)\frac{\lambda^n}{n!}T_2(n) &= \lambda^2 \\
T_2(n) &= \frac{\partial^n(\exp(\lambda)\lambda^2)}{\partial n^n} \\
T_2(0) &= 0 \\
T_2(1) &= 0 \\
T_2(n) &= n(n-1).
\end{aligned}$$

There is a clear "contradiction" between the two estimates.

2.2.2 Estimation principle: Maximum Likelihood

The principle amoubts to estimating parameters so that the observed sample be the most likely with this set of parameters. It is a principle, nothing proves *a priori* that it has good properties. The likelihood function can be defined as

$$L(\theta \mid x) = f(x \mid \theta),$$

where f is either the density function or the probability mass function. The standard example is a normal population, and an i.i.d. sample.

$$f(X|\ \boldsymbol{\mu}, \boldsymbol{\sigma}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2\sigma^2}\right].$$

Let us apply the ML principle, and maximize the likelihood function:

$$\max_{m,s} L(m, \mathbf{s} \mid \mathbf{x}) = (2\pi s^2)^{-\frac{n}{2}} \exp\left[-\frac{\sum_{i=1}^{n}(x_i-m)^2}{2s^2}\right]$$

Usually we use the logarithm of the likelihood function. Logarithm is a monotone transformation, therefore the maximum of the log-likelihood function is the same as that of the likelihood function. Interesting statements refer, in general, to the log-likelihood (see below, for instance, the Cramér-Rao inequality.)

Proposition:

$$m^{ML} = \overline{X}_n$$

$$s^{2ML} = s_u^2 = \frac{1}{n}\sum(X_i - \overline{X_n})^2.$$

This example shows that an ML estimator is not necessarily unbiased.

2.2.3 Why is ML sensible?

ML has the attractive Invariance property: if t is an ML estimate for θ, then $r(t)$ is an ML estimate of $r(\theta)$ for any r function.

For example, If we have an unbiased estimate of a parameter then the square (or square root) of the estimate is not unbiased for the square of the parameter.

If an ML estimate is unbiased then it is efficient in the sense that it has minimum variance among the set of unbiased estimators.

Cramér-Rao inequality Let

$$I(\theta) = -E(\frac{\partial^2 \log L(\theta)}{\partial \theta^2})$$

be the Fisher-information.

Then the variance of any unbiased estimator is as large as the inverse of the Fisher.information $\frac{1}{I(\theta)}$.

Definition: An estimator is called consistent for θ, if

$$p\lim_n T_n(X_n) = \widehat{\theta}.$$

In general in i.i.d. samples sample moments are consistent estimates of the theoretical moments if their second moment exists. This can be derived from the Law of Large Numbers.

Asymptotic properties of ML estimators (under some regularity conditions) It can be proved that

1. ML estimators are consistent.
2. ML estimators are asymptotically normal, in the sense that

$$\sqrt{n}(T_n(X_n) - \widehat{\theta})$$

converges to a normal distribution.

3. The asymptotic variance of ML estimators achieves the Cramér-Rao lower bound among consistent estimators.

2.3 Second approach: hypothesis testing

Here we postulate something about the true θ, and either accept or reject this hypothesis. A testing procedure is a rule: if $T(x)$ is an element of Ω_0 the hypotheses is accepted, if not, it is rejected. The null-hypothesis can be identified with $\theta \in \Theta_0$, while the alternative with $\theta \in subset(\Theta - \Theta_0)$. The $\theta \in \Theta$ assumption is sometimes called the maintained hypothesis.

A Type 1 error occurs if we we reject the null, though it is true. A Type 2 error occurs if we accept the null, though it is false. As the alternative hypothesis contains many possible truths in general, there is no single Type 2 error, for each true parameter there belongs one. The size of the test is the the probability of the Type 1 error, while the power of the test is 1 minus the probabiity of the Type 2 error, thus the power is a function of the true parameter.

Analogously to the ML principle we have a general testing principle: the likelihood ratio (LR) test.

The LR test statistic for $\theta \in \Theta_0$ is defined as

$$\lambda = \frac{\sup_{\theta \in \Theta_0} L(\theta, x)}{\sup_{\theta \in \Theta - \Theta_0} L(\theta, x)}.$$

An LR test is any procedure with rejection region

$$\lambda(x) \leq c, 0 \leq c \leq 1.$$

It can be proved that it is a uniformly most powerful (UMP) test: for each size (i.e. type-1 error) it has the highest power.

An example: mean with known variance (the normal case) The null hypothesis is $\mu = \mu_0$, and $\sigma = \sigma_0$ known, and the level is α. We know that $z = \sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma_0}$ is standard normal, if the null hypothesis is true. Then

$$P(abs(\sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma_0}) > z_{\alpha/2}) = \alpha$$

determines $z_{\alpha/2}$. Here for $z = \sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma_0}$:

$$P\left[-z_{\alpha/2} \leq z \leq z_{\alpha/2}\right] = 1 - \alpha,$$

Therefore if

$$abs(\sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma_0}) \leq z_{\alpha/2},$$

the null is accepted, otherwise it is rejected. In an alternative formulation the acceptance region is

$$\overline{X}_n \in (\mu_0 - \frac{\sigma_0 z_{\alpha/2}}{\sqrt{n}}, \mu_0 + \frac{\sigma_0 z_{\alpha/2}}{\sqrt{n}}).$$

Definition: Pivotal quantity: a statistics is a pivotal quantity if its distribution does not depend on θ. For example $z = \sqrt{n}\frac{\overline{X}_n - \mu_0}{\sigma_0}$ is pivotal in the former example. If the variance is unknown

$$t = \sqrt{n}\frac{\overline{X}_n - \mu}{s}.$$

is pivotal. Pivots help in finding tests.

Testing paradox There is an experiment, and a test procedure with size 0.05. Suppose it defines an acceptance region of (-2,2). The test statistics turns out to be 1.9, thus the null is accepted. Later on the investigator discovers that 2 could never be reached as the measuring device had limits. On the other hand the limits were never hit during the experiment. Still the statistician should recalculate the acceptance region, as the "true" probability of the original acceptance region is higher than 0.95, and after narrowing the acceptance region, it may not contain 1.9.

Is it reasonable that the outcome of a test depend on evidence that did not "materialize"?

2.4 Third approach: interval estimation

Let us determine an interval that will probably "cover" the true parameter. This is defined by a pair of statistics with $S_L(x) < S_U(x)$.

Example: normal population with known σ We know that

$$z = \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}$$

is standard normal. Let $1-\alpha$ be the confidence level, and F the standard normal cdf. $z_{\alpha/2}$ is implicitly defined from

$$F(z_{\alpha/2}) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{z^\alpha} \exp(-\frac{x^2}{2})dx = 1 - \frac{\alpha}{2}.$$

Then

$$P\left[-z_{\alpha/2} \leq \sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \leq z_{\alpha/2}\right] = 1-\alpha.$$

therefore

$$P\left[\overline{X}_n - \frac{1}{\sqrt{n}}\sigma z_{\alpha/2} \leq \mu \leq \overline{X}_n + \frac{1}{\sqrt{n}}\sigma z_{\alpha/2}\right] = 1-\alpha.$$

Proposition: If

$$\begin{aligned} T_2 &= \overline{X}_n + \frac{1}{\sqrt{n}}\sigma z_{\alpha/2} \\ T_1 &= \overline{X}_n - \frac{1}{\sqrt{n}}\sigma z_{\alpha/2} \end{aligned}$$

then with probability $1-\alpha$ μ is included in the (T_1, T_2) random interval. Notice what is random: the endpoints of the interval, but not the parameter. This confidence interval is created by "inverting" the corresponding test.

Confidence interval paradox Let x be uniform on $(\theta - 1, \theta + 1)$. Suppose we have an independent sample of 4 observations. Then define $x_{\min} = \min(x_i), x_{\max} = \max(x_i)$. Then

$$\begin{aligned} P(x_{\min} &> \theta) = 1/16 \\ P(\theta &> x_{\max}) = 1/16 \end{aligned}$$

and

$$P(x_{\min} < \theta < x_{\max}) = 7/8.$$

Thus $(x_{\min}, x_{\max})$ is a confidence interval of size 7/8.
Suppose that in the sample

$$\begin{aligned} x_{\min} &= 1.5 \\ x_{\max} &= 2.7 \end{aligned}$$

values are realized. Then we KNOW for sure that θ must be between $x_{\min}$ and $x_{\max}$. The lesson is that certain realizations provide "better" information on the unknown parameter than others.

2.5 Bayesian-statistics

As the above "paradoxical" examples show the classical estimation and testing procedures may have strange implications in certain cases. Many statisticians have been interested in general principles, that would be helpful to prove the reasonability of statistical procedures.

A large part of statistics amount to data compression, i.e. when the dimension of a T statistic is (much) smaller than the dimension of X. Data compression can be approached from other perspectives as well, but classical statisticians considered it as a problem in probability theory. Historically the most important concept is sufficient statistic. Intuitively if we have a sufficient statistic then we do not need anything else for estimation purposes.

Let

$$F(X \mid \theta)$$

the sampling distribution for given θ. $S(X)$ is called sufficient for θ if the conditional distribution of X with condition S

$$G(X \mid S(X))$$

is independent of θ.

Theorem 1 *Let $H(S(X) \mid \theta))$ the distribution of S. S is sufficient for θ, if*

$$\frac{F(X \mid \theta)}{H(S(X) \mid \theta))}$$

is constant as a function of θ.

This theorem provides a method to derive sufficient statistics for specific cases. In a sense if we have a sufficient statistic then we should not search further, we have a method to summarize the data without loss of information. For instance if we have a sample of independent characteristic variables, where the only unknown parameter is p (the probability of 1) then the sum of the variables (which has a binomial distribution) is a sufficent statistic. Intuitively we should not care about the exact sequence of 0s and 1s, it is enough to count them.

Many believe that the so-called Sufficiency Principle is an axiom that sound statistical procedures must satisfy. This Principle asserts that two experimental results that result in the the same sufficient statistic must provide the same evidence. A second axiom whose plausibility seems obvious is the Conditionality Principle: Suppose that two experiments with the same parameter space are randomized with equal probability. The eventually performed experiment must have the same evidence as the same experiment performed without the randomization.

With respect to these two principles the Likelihood Principle appears to be not so obvious. It asserts that if two experiments have the same likelihood function then all evidence derived from them must be the same. However, the celebrated Birnbaum Theorem states that the Sufficiency Principle and the Conditionality Principle are equivalent with the Likelihood Principle.

It can be proved, however, that classical testing and confidence interval formation procedures do not satisfy the Likelihood Principle. Thus classical statistics does not satisfy either the Conditionality Principle or the Sufficiency Principle.This argument justifies a search for a different outlook for statistics, which is provided by the Bayesian approach.

2.5.1 Bayesian principles and concepts

There is an important addition to the basic classical model. Bayesians equip the parameter space with a prior distribution for the parameters: $p(\theta)$. Then if the conditional distribution of the sample for any parameter $f(x \mid \theta)$ is given then the posterior distribution of the parameters can be derived from Bayes' Theorem.

Proposition 2 *Bayes' Theorem*

$$p(\theta \mid x) = \frac{f(x \mid \theta)p(\theta)}{\int\limits_{\Theta} f(x \mid \theta)p(\theta)d\theta'}.$$

To simplify derivations we can observe that the posterior is proportional to the product of the conditional and the prior distibutions:

$$p(\theta \mid x) \propto f(x \mid \theta)p(\theta).$$

Bayesian updating implies

$$p(\theta \mid x_1, x_2) \propto f(x_1, x_2 \mid \theta) p(\theta).$$

$$p(\theta \mid x_1, x_2) \propto f(x_2 \mid x_1, \theta) p(\theta \mid x_1).$$

Thus new information can be accommodated recursively, each piece of new data will cause an update of the posterior, the essential goal of the Bayesian analysis.

Obviously in practice there remains the question of selecting the prior. If one would like to obtain analytical results the choice must be fine-tuned. For an example let us consider again n observations of a characteristic variable where $P(1) = \theta$. The likelihood function (the conditional distribution) is binomial:

$$p(x \mid \theta) = \theta^{\sum x_i} (1-\theta)^{n - \sum x_i}$$

We look for conjugate priors, where for a given prior and likelihood the posterior belongs to the same family as the prior. For instance the binomial likelihood and beta prior are conjugate pairs, where the beta distribution is defined as

$$p(\theta) = \frac{\Gamma(\alpha) + \Gamma(\beta)}{\Gamma(\alpha + \beta)} \theta^{\alpha - 1} (1-\theta)^{\beta - 1},$$

implying

$$\begin{aligned} E(\theta) &= \frac{\alpha}{\alpha + \beta} \\ var(\theta) &= \frac{\alpha\beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}. \end{aligned}$$

Then the posterior is beta with:

$$\begin{aligned} \alpha_1 &= \alpha + \sum x_i \\ \beta_1 &= \beta + n - \sum x_i. \end{aligned}$$

From this it is easy to see that

$$E(\theta \mid x) = \left(\frac{\alpha + \beta}{\alpha + \beta + n} \right) \frac{\alpha}{\alpha + \beta} + \left(\frac{n}{\alpha + \beta + n} \right) \frac{1}{n} \sum x_i,$$

thus the posterior's expected value of the parameter is a weighted average of the prior mean and the sample mean. One can see that as $n \to \infty$ the prior becomes inessential. So if we take this posterior expected value as the "estimate" of θ then this would converge to the ML estimate as the sample size increases.

Some general statements can be proved about Bayesian inference for large n

(a) The role of the prior gets smaller and smaller.

(b) The posterior converges to a degenerate distribution.

(c) The posterior is asymptotically normal with mean the "true" θ.

So in the infinite limit Bayesian and classical inference may not be so different. But in the middle range how could we define a Bayesian estimate?

2.5.2 Bayesian point estimation and loss function optimality

For Bayesians the "truth" is embodied in the posterior distribution of the parameter, solely. Still it is considered rightful to give a point estimate, for some purpose. However, the purpose must be defined precisely.

If we want to use a point estimate or some purpose we have to know what the costs of making mistakes are. We define a loss function as

$$L(\theta, \widehat{\theta}),$$

where θ is the true parameter, and $\widehat{\theta}$ is the would-be estimate based on the posterior. It is not necessary but customary to have

$$L(\theta, \widehat{\theta}) = 0, \theta = \widehat{\theta}.$$

Then the expected loss is defined as

$$R(\theta, d(X)) = E_\theta L(\theta, d(X)),$$

where $d(X)$ is some point estimator, and the expectation is taken according to the posterior. Obviously two $d()$-s may not be ordered by their expected losses unequivocally, for different x observations one or the other can be more efficient.

The Bayes-rule for point estimation is defined as

$$d^B(X) = \arg\min_{d(X)} \int_\Theta R(\theta, d(X)) p(\theta) d\theta.$$

As

$$\int_\Theta (\int_X L(\theta, d(x)) f(X \quad | \quad \theta) dx) p(\theta) d\theta =$$
$$\int_X (\int_\Theta L(\theta, d(x)) p(\theta \quad | \quad x) d\theta) m(x) dx.$$

minimization is equivalent to minimizing the posterior expected loss $\int_\Theta L(\theta, d(X)) p(\theta \mid X) d\theta)$ for each x, a mechanical procedure to find the Bayesian estimate for a given loss function. The Bayes-rule as a statistical procedure satisfies the Likelihood Principle.

Proposition 3 *For a quadratic loss function the Bayes rule requires:* $d(x) = E(\theta \mid x)$*. For absolute error loss* $d(x)$ *must be the median of* $p(\theta \mid x)$*.*

Thus the estimate of θ as the posterior expected value in the example above can be justified with having a quadratic loss funcion in mind.

2.5.3 Bayesian interval estimation

For some parameter θ and an α credibility level we look for a credibility interval as

$$P(\theta_A \leq \theta \leq \theta_B) = \alpha.$$

If there are many, we choose the shortest. Again this procedure satisfies the Likelihood Principle, and is based only on observed, and not on would-be, data.

2.5.4 Bayesian testing

Bayesian testing is esssentially model comparison. Suppose there exist two possible models:

Model 1 with $p_1(\theta_1), f_1((x \mid \theta_1)$ and Model 2 with $p_2(\theta_2), f_2((x \mid \theta_2)$.The marginal posterior likelihoods are:

$$f_i(x) = \int_{\Theta} f_i(x \mid \theta_i) p_i(\theta_i) d\theta_i'.$$

The idea is to compare the marginal posterior likelihoods: $f_1(x), f_2(x)$. One can calculate the Bayes factor:

$$\frac{f_1(x)}{f_2(x)}.$$

If it is larger than 1 then we can say that data are in better accordance with Model 1, than with Model 2. Another possible quantity to calculate is

$$\frac{p_1(M_1) f_1(x)}{p_2(M_2) f_2(x)},$$

where we give, in true Bayesian spirit, a prior chance to both Model 1 and Model 2.

2.5.5 Practical considerations

Bayesian inference at first sight is a mechanical procedure: set the prior and the likelihood, and derive the posterior after data arrive. However, calculating a distribution may not be done analytically even if the prior and the likelihood are analytic. And in the next step the posterior becomes the prior which may be non-analytical from now on. Then to compute point estimates we need to find

marginal distributions and moments from the posterior. Fortunately numerical integration or simulation can help us to carry out our plan. However, these may require a lot of calculation. This is partly the reason why the increase in the efficiency of computation technology gave an important shove to Bayesian statistics, formerly researchers were largely constrained to look for appropriate conjugate priors.

2.6 Literature

Casella, G., & Berger, R. L. (2002). Statistical inference (Vol. 2). Pacific Grove, CA: Duxbury.

Greenberg, E. (2012). Introduction to Bayesian econometrics. Cambridge University Press.

Samaniego, F. J. (2010). A comparison of the Bayesian and frequentist approaches to estimation. Springer Science & Business Media.

3 Classical conditional estimation: regression

3.1 Introducing the regression problem

We observe the heights of n Englishmen, and the heights of their parents. Knowing the parents' heights, what would be our best guess for the vertical extension of any Englishman (not just those in the sample)? This was roughly Galton's original regression problem.

The problem can be generalized as follows. Given observations on y (response), and $\mathbf{X}$ (features, explanatory variables) establish some relationship that could predict (explain, describe) y based on information on $\mathbf{X}$.

Several approaches exist for solving this (ill-defined) problem, the traditional statistical approach is based on probability theory. Here one assumes that there exists an underlying probability measure that describes the "population" in question. Then, we make assumptions about this "theoretical" population, we define our estimands, i.e. certain unknown properties of the population distribution. Next, a method of sampling is determined (or assumed if data are given), where identically independently distributed (i.i.d.) samples are typical.

What are the reasons for assuming a probability structure? Usually even for the same Xs we find different ys, therefore a deterministic functions do not conform to the facts, in general.

Let us start with discussing some population concepts (probability theory), then we proceed to sample concepts (statistics).

3.2 Conditional expectations and population regression

We start by operationalizing the concept of "best guess"! We are looking for a function $f(\mathbf{x})$ that minimizes

$$E(y - f(\mathbf{x}))^2,$$

the Mean Squared Error (MSE).

Here we must introduce certain important concepts: conditional mean or expectation of a random variable, its properties and rules of operation.

3.2.1 Properties of the conditional expectation

Let y be a random variable, and $\mathbf{x}$ a random vector.

Proposition 4 *The Law of Iterated Expectations*

$$E(y) = E(E(y \mid \mathbf{x}))$$

The unconditional expectation is the expectation of the conditional expectations. This is an important theorem, that is frequently used in theoretical derivations.

Proposition 5 *CEF (conditional expectation function) decomposition*

There exists ϵ, for which

$$\begin{aligned} y &= E(y \mid \mathbf{x}) + \epsilon \\ E(\epsilon &\mid \mathbf{x}) = 0, \end{aligned}$$

and ϵ is uncorrelated with any function $h(\mathbf{x})$, as $E(h(\mathbf{x})\epsilon) = E(E(h(\mathbf{x})\epsilon \mid \mathbf{x})) = E(h(\mathbf{x})E(\epsilon \mid \mathbf{x}))$, by the Law of Iterated Expectations. It is a basic property, that can be used again and again in proving theorems.

Proposition 6 *CEF as the best predictor*

$$E(y \mid \mathbf{x}) = \arg\min_{m(\mathbf{x})}(E(y - m(\mathbf{x}))^2)$$

where $m(\mathbf{x})$ is any function.
Proof:

$$\begin{aligned} E(y - m(\mathbf{x}))^2 &= E(y - E(y \mid \mathbf{x}))^2 + 2E(y - E(y \mid \mathbf{x})(E(y \mid \mathbf{x}) - m(\mathbf{x})) \\ &+(E(y \mid \mathbf{x}) - m(\mathbf{x}))^2. \end{aligned}$$

The first term is irrelevant, and the middle term is 0 by the CEF decomposition.

This theorem is important for practical work, as it asserts that if we look for an estimates that are best in the MSE sense, then our natural candidate is to have an estimate for the CEF.

3.2.2 Linear projection (population regression)

It may be difficult to obtain the CEF even theoretically. Let us look for a simpler (parametric) solution!

Find the affine transform of the x variables that minimizes the MSE:

$$\min_{\beta}(E(y - \alpha - \mathbf{x}'\boldsymbol{\beta})^2).$$

The first order conditions are

$$\begin{aligned} E(x_j(y - \mathbf{x}'\boldsymbol{\beta})) &= 0 \\ E((y - \mathbf{x}'\boldsymbol{\beta})) &= 0. \end{aligned}$$

Define the residual variable as:

$$\epsilon = y - \mathbf{x}'\boldsymbol{\beta}.$$

Then the first order conditions assert that

$$E(x_j\epsilon) = 0,$$

for all j, i.e. the residual is orthogonal to (uncorrelated with) the x_j variables. Also it is true that

$$E(\epsilon) = 0.$$

By analogy, it is called linear projection, or, alternatively, the population regression.

From now on we assume that a constant (a degenerate random variable) belongs to $\mathbf{x}$, and we can dispose of α (the constant). Alternatively we could assume that each variable is replaced by its centralized counterpart, that is the mean is subtracted.

The solution can be written compactly as:

$$\boldsymbol{\beta} = E((\mathbf{x}\mathbf{x}')^{-1})E(\mathbf{x}y).$$

The two ϵs (defined by the CEF and the linear projection, respectively), may be different. Notice that $E(\epsilon \mid \mathbf{x}) = 0$ is not necessarily fulfilled by the regression residual.

Proposition 7 *Regression parameters*

$$\boldsymbol{\beta}_j = \frac{cov(y, \widetilde{x_j})}{var(\widetilde{x_j})}$$

where $\widetilde{x_j}$

$$\widetilde{x_j} = x_j - \sum_{i \neq j} \boldsymbol{\beta}_{ji} x_i,$$

and β_{ji} are parameters of the projection of x_j on $x_1, .x_{j-1}.x_{j+1}.x_n$.
Proof:

$$y = \sum_{j=1}^{k} \boldsymbol{\beta}_j x_j + \varepsilon,$$

Multiply with $\widetilde{x_j}$, and take expectations. As

$$\begin{aligned} cov(x_i, \widetilde{x_j}) &= 0, j \neq i \\ cov(\varepsilon, \widetilde{x_j}) &= 0 \\ cov(x_j, \widetilde{x_j}) &= var(\widetilde{x_j}) \end{aligned}$$

it follows that

$$cov(y, \widetilde{x_j}) = \boldsymbol{\beta}_j var(\widetilde{x_j}).$$

It is also true that

$$\boldsymbol{\beta}_j = \frac{cov(\widetilde{y_{-j}}, \widetilde{x_j})}{var(\widetilde{x_j})}$$

where $\widetilde{y_{-j}}$ is the "residual"from the projection of y on the $x_k - s$ (except x_i).

$$\widetilde{y_{-j}} = y - \sum_{i \neq j} \boldsymbol{\beta}_{ji} x_i,$$

On the other hand

$$\boldsymbol{\beta}_j = \frac{cov(\widetilde{y_{-j}}, x_j)}{var(x_j)}$$

only if x_j is orthogonal to the others.

Proposition 8 *the relationhip between regression and linear CEF*

If the conditional expectation function is linear then it coincides with the linear projection.

Proof: $\mathbf{x}$ is uncorrelated with the decomposition error of the conditional expectation.

One celebrated case when this is necessarily true is when the variables are jointly normal.

Proposition 9 *Best linear prediction (least squares problem)*

The linear projection is the best linear predictor of y (in the MMSE (minimum mean squared error) sense).

In other words $\boldsymbol{\beta}$ solves

$$\min E((E(y \mid \mathbf{x}) - \mathbf{x}'.$$

Proposition 10 *The linear projection (population regression) is the minimum MSE linear approximation to the CEF. In other words, $\boldsymbol{\beta}$ solves:*

$$\min_{\beta} E\left[\left(E(y \mid \mathbf{x}) - \mathbf{x}'\boldsymbol{\beta}\right)^2\right].$$

Proof:

$$\begin{aligned} (y - \mathbf{x}'\boldsymbol{\beta})^2 &= (y - (E(y \mid \mathbf{x}))^2 + \\ 2(y - (E(y &\mid \mathbf{x}))(E(y \mid \mathbf{x}) - \mathbf{x}'\boldsymbol{\beta}) + \\ (E(y &\mid \mathbf{x}) - \mathbf{x}'\boldsymbol{\beta})^2. \end{aligned}$$

The first term is irrelevant, the second term is 0, by the CEF decomposition property, thus this problem has the same solution as the least squares problem.

This theorem has great practical importance as it suggests that in short of being able to estimate the CEF we may avail ourselves with estimating the population regression. However, as the examples below show, this standpoint may be overoptimistic.

Examples Suppose that x and z are independent standard normal variables. If x is standard normal, then x^2 is χ^2 with 1 degree of freedom.

Example 1 Project $y = 3x^2 + z$ on x

$$\begin{aligned} E(x(3x^2 + z - a - bx)) &= 0, \\ E(3x^2 + z - a - bx) &= 0 \\ a &= 3, b = 0 \\ proj(y \mid x) &= 3. \end{aligned}$$

However the CEF is clearly:

$$E(y \mid x) = 3x^2.$$

In this case the projection is very poor approximation to the CEF.

Example 2 Now project $y = 3x^2 + z$ on z and a constant.

$$\begin{aligned} E(z(3x^2 + z - a - cz)) &= 0 \\ E(3x^2 + z - a - cz) &= 0 \\ a &= 3, c = 1 \\ proj(y \mid z) &= 3 + z. \end{aligned}$$

The CEF now is exactly the same:

$$E(y \mid z) = 3 + z.$$

Example 3 Now project $y = 3x + z$ on x.

$$\begin{aligned} E(x(3x + z - a - bx)) &= 0 \\ E(3x + z - a - bx) &= 0 \\ a &= 0, b = 3 \\ proj(y \mid x) &= 3x. \end{aligned}$$

The CEF:

$$E(y \mid x) = 3x.$$

There is, again, equivalence.

Example 4 Now project $y = 3x + xz$ on x, and a constant.

$$\begin{aligned} E(x(3x + xz - a - bx)) &= 0 \\ E(3x + xz - a - bz) &= 0 \\ a &= 0, b = 3 \\ proj(y \quad &| \quad x, 1) = 3x. \end{aligned}$$

The CEF:

$$E(y \mid x,) = 3x.$$

The linear projection in partitioned form We are frequently interested in finding out what effects the inclusion or exclusion of some variables would imply. Writing down the regression in paritioned form help in this.

$$\begin{matrix} E(x_1 y) \\ E(x_2 y) \end{matrix} = E \begin{pmatrix} \mathbf{x}_1 x_1^{'} & \mathbf{x}_1 x_2^{'} \\ \mathbf{x}_2 \mathbf{x}_1' & \mathbf{x}_2 x_2^{'} \end{pmatrix} \begin{matrix} \beta_1 \\ \beta_2 \end{matrix}.$$

It can be derived that

$$\boldsymbol{\beta}_1 = \mathbf{E}(\mathbf{x}_1 \mathbf{x}_1^{'})^{-1} E(\mathbf{x}_1 (\mathbf{y} - \mathbf{x}_2' \boldsymbol{\beta}_2)).$$

Then let

$$\boldsymbol{\beta}_1^s = \mathbf{E}(\mathbf{x}_1 \mathbf{x}_1^{'})^{-1} E(\mathbf{x}_1 \mathbf{y}).$$

This way we obtain the omitted variable formula:

$$\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^s - E(\mathbf{x}_1 \mathbf{x}_1^{'})^{-1} E(\mathbf{x}_1 \mathbf{x}_2^{'}) \boldsymbol{\beta}_2.$$

This gives the change in the regression coefficients of the first group of variables due to the omission of the second group of variables.

With two variables the formula is clearer. Consider the following linear projection:

$$y = \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where

$$E(x_i) = 0, E(x_i \epsilon) = 0, i = 1, 2,$$

Then

$$\begin{aligned} cov(y, x_1) &= \beta_1 var(x_1) + \beta_2 cov(x_1 x_2) \\ \beta_1 &= \frac{cov(y, x_1)}{var(x_1)} - \frac{cov(x_1 x_2)}{var(x_1)} \beta_2. \end{aligned}$$

But

$$proj(y \mid x_1) = \beta_1^s x_1 = \frac{cov(y, x_1)}{var(x_1)}$$

and

$$proj(y \mid x_1) = \beta_{12} x_1 = \frac{cov(x_1, x_2)}{var(x_1)}$$

therefore

$$\beta_1^s = \beta_1 + \beta_{12}\beta_2.$$

With some abuse of words it is frequently interpreted that the full effect of x_1 on y equals the sum of the direct effect and the indirect effect via x_2. It must be clear that talking of effects is totally unwarranted if the word "effect" is used in the normal sense. Later in these notes we will consider cases when this language is justified.

3.3 The classical statistical approach to regression

There is a finite (of size n) i.i.d. sample from a population, with the ith observation $(y_i, x_{1i}, ...x_{ki})$. We want to estimate the estimands (some parameters of this sample) in order to give a good guess of y. The key idea is to substitute sample moments for theoretical moments.

For instance:

$$E(x_j y) \cong \frac{1}{n} \sum_i x_{ij} y_i.$$

The Law of Large numbers says that this is correct with high probability if n is large.

3.3.1 What is a linear (sample) regression?

Here we "imitate" the population regression.

OLS (ordinary least squares) principle OLS is an estimation principle, applied to our problem it requires that we minimize

$$\sum_i (y_i - \sum_j x_{ij} b_j)^2.$$

This minimization problem is the 'sample" equivalent of the theoretical projection problem.

By the first order conditions of the minimum this leads to the following normal equations

$$\sum_i (y_i - \sum_j x_{ij} b_j) x_{ik} = 0, \forall k.$$

These are called the sample moment orthogonality conditions. Therefore it is also called a method of moments estimator.

In matrix form the normal equations can be written as

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = 0.$$

From this the explicit solution follows:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}^{'}\mathbf{y}.$$

The sample equivalent of the omitted variable formula can be derived as follows. Let us write the sample regression in partitioned form

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon},$$

$$\mathbf{y} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{bmatrix} + \boldsymbol{\epsilon}$$

where $\mathbf{X}_1$ nxk_1 and $\mathbf{X}_2$ $nx(k-k_1)$.

$$\begin{matrix}\mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y}\end{matrix} = \begin{matrix}\mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 & \mathbf{b}_1 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 & \mathbf{b}_2\end{matrix}.$$

Then

$$\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{y} - \mathbf{X}_2\mathbf{b}_2).$$

the "long" parameter vector.

Let

$$\mathbf{b}_1^s = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$$

be the "short" parameter vector.

then

$$\mathbf{b}_1 = \mathbf{b}_1^s - (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\mathbf{b}_2.$$

The omitted variable formula in matrix notation is

$$\mathbf{B}_{12} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2.$$

$$\mathbf{b}_1^s = \mathbf{b}_1 + \mathbf{B}_{12}\mathbf{b}_2.$$

Again by some abuse of words one say that $\mathbf{B}_{12}\mathbf{b}_2$ measures the effect of omitting $\mathbf{x}_2$ **.** If $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$ or $\mathbf{b}_2 = 0$ then the long coefficients equal the short ones.

OLS properties

Proposition 11 $\mathbf{b}_{ols}$ *is an unbiased estimator of* $\boldsymbol{\beta}$*, moreover it is a BLUE (minimal variance unbiased linear) estimator. If* $lim_{n\to\infty}((\mathbf{X}'\mathbf{X})/n) = Q$*, then it is consistent, too.*

Proof:
Unbiasedness:

$$\begin{aligned}
\mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}^{'}\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}^{'}(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}) \\
\mathbf{b} &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}^{'}\boldsymbol{\epsilon} \\
E(\mathbf{b}) &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}^{'}E(\epsilon) = \boldsymbol{\beta}
\end{aligned}$$

BLUE:

$$Var(\mathbf{b}) = E(((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})-\boldsymbol{\beta})((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})-\boldsymbol{\beta})'$$

$$\begin{aligned}
&= E(((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon})((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon})' \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^{'}\mathbf{X})^{-1}
\end{aligned}$$

Let $\mathbf{b}^c$ be another unbiased linear estimator:

$$\mathbf{b}^c = \mathbf{b} + \mathbf{C}\mathbf{y}.$$

$$\mathbf{b}^c = ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{C})(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon})$$

If $\mathbf{b}^c$ is unbiased then $\mathbf{C}\mathbf{X} = \mathbf{0}$ and

$$\begin{aligned}
Var(\mathbf{b}^c) &= E((\mathbf{b}^c-\boldsymbol{\beta})(\mathbf{b}^c-\boldsymbol{\beta})' \\
E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{C})\boldsymbol{\epsilon}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{C})\boldsymbol{\epsilon})') &= \\
&\quad \sigma^2((\mathbf{X}'\mathbf{X})^{-1} + \mathbf{C}\mathbf{C}').
\end{aligned}$$

$$Var(\mathbf{b}^c) = var(\mathbf{b}) + \mathbf{Q},$$

where $\mathbf{Q}$ positive semidefinite.
Consistency:

$$\begin{aligned}
\mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}^{'}\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}^{'}(\mathbf{X}\boldsymbol{\beta}+\epsilon) \\
\mathbf{b} &= \boldsymbol{\beta} + n(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}^{'}\frac{\boldsymbol{\epsilon}}{n} \\
\operatorname*{Pr\,lim}_{n\to\infty}\mathbf{b} &= \boldsymbol{\beta} + \mathbf{Q}^{-1}\operatorname*{Pr\,lim}_{n\to\infty}\frac{\mathbf{X}^{'}\boldsymbol{\epsilon}}{n} = \boldsymbol{\beta}
\end{aligned}$$

(Because $\mathbf{X}'E(\frac{\boldsymbol{\epsilon}}{n}) = 0$, and $\lim_{n\to\infty} \mathbf{X}'var\frac{\boldsymbol{\epsilon}}{n} = 0$, (Law of Large Numbers).)

This Theorem makes statements conditional on $\mathbf{X}$. There are two possible readings: 1. OLS is the best estimate of the linear projection parameters for given $\mathbf{X}$. 2. If the CEF is linear then OLS is the best linear estimate of the CEF. For a given $\mathbf{X}$ the distinctive properties of the two types of residuals do not matter.

Estimation of the variance Let

$$\mathbf{u} = \mathbf{y} - \mathbf{X}\mathbf{b}$$

be the OLS residual.
Then:

$$\mathbf{X}'\mathbf{u} = \mathbf{0}$$

as

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{0}.$$

In particular:

$$\mathbf{1}'\mathbf{u} = \mathbf{0},$$

if a constant appears in the regression.
Let

$$s^2 = \frac{1}{n-k}\sum u_i^2 = \frac{\mathbf{u}'\mathbf{u}}{n-k}.$$

(where s is called the standard error of the regression). Then

$$E(s^2) = \sigma^2,$$

and therefore $s^2(\mathbf{X}'\mathbf{X})^{-1}$ is an unbiased estimate of $Var(\mathbf{b})$.
Proof:

$$\begin{aligned} \mathbf{u} = \mathbf{y} - \mathbf{X}\mathbf{b} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \\ \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) &= \\ \boldsymbol{\epsilon} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} &= \\ \mathbf{M}\boldsymbol{\epsilon} \end{aligned}$$

where

$$\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

$$\mathbf{M} = \mathbf{M}^2.$$

Therefore:

$$\mathbf{u}'\mathbf{u} = \boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon}.$$

By the properties of idempotent matrices:

$$\mathbf{E}(\mathbf{u}'\mathbf{u}) = \mathbf{E}(\boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon}) = \mathbf{E}(tr(\boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon})) = \mathbf{E}(tr(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{M})) = tr(\sigma^2\mathbf{I}\mathbf{M}) = \sigma^2 tr(\mathbf{M}).$$

As

$$\mathbf{N} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

is idempotent, too, and as $\mathbf{M} = \mathbf{I} - \mathbf{N}$, $rank(I) = n$ and $rank(N) = k$

$$tr(\mathbf{M}) = n - k.$$

Moreover:

$$\mathbf{M}\mathbf{N} = \mathbf{0}.$$

Thus:

$$\sigma^2 = E(\frac{\mathbf{u}'\mathbf{u}}{n-k}) = E(s^2).$$

Sometimes the question is raised whether it is wise to use any available variables in a regression. The next statement shows why this can be disadvantageous.

We call the second set of variables redundant, if $\mathbf{b}_2 = 0$.

Then

$$\mathbf{cov}(\mathbf{b}_1^s) = \sigma^2(\mathbf{X}_1'\mathbf{X}_1)^{-1}$$

$$\begin{aligned} \mathbf{N}_2 &= \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2' \\ \mathbf{M}_2 &= \mathbf{I} - \mathbf{N}_2. \end{aligned}$$

$$cov(\mathbf{b}_1) = \sigma^2(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}$$

and

$$cov(\mathbf{b}_1^s)^{-1} - \mathbf{cov}(\mathbf{b}_1)^{-1} = \sigma^2(\mathbf{X}_1'\mathbf{N}_2\mathbf{X}_1),$$

is positive definite. In other words redundant variables reduce the precision of our estimates.

3.3.2 Normality and ML

Suppose now that $\varepsilon_i \sim N(0, \sigma^2)$ $i = 1, ...n$. Then one can write the likelihood function as

$$L(\mathbf{y}, \mathbf{X}; b, \mathbf{s}^2) = \sigma^{-n}(2\pi)^{-\frac{n}{2}} \exp(-\frac{\sum_{i=1}^{n}(y_i - \mathbf{X}_i'\boldsymbol{\beta})^2}{2\sigma^2}).$$

and the log-likelihood as:

$$\ln L = -n \ln s - n \ln(\sqrt{2\pi}) - \frac{1}{2s^2}\sum(y_i - \mathbf{X}_i'b)^2.$$

The ML Principle requires choosing $(\mathbf{b}_{ML}, s^2_{ML})$ so that the likelihood function be maximized.

The ML estimator for β is the same as the OLS. The first order condition for s^2 is

$$-\frac{n}{s_{ML}} + \frac{1}{s^3_{ML}}\sum(y_i - X_i'b_{ML})^2 = 0.$$

From which

$$s^2_{ML} = \frac{ESS}{n},$$

which is biased downwards, as we know. However for the modified (unbiased) estimate

$$s^2 = \frac{n}{n-1}s^2_{ML}.$$

It can be proved that

$$(n-k)s^2/\sigma^2 \sim \chi^2_{n-k}.$$

Proof:

$$\frac{\varepsilon'}{\sigma}\mathbf{M}\frac{\varepsilon}{\sigma} = \frac{\mathbf{u}'\mathbf{u}}{\sigma^2} = \frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{n-k}.$$

This statement is important for deriving test statistics. In particular, from this it turns out that the elements of $s\sqrt{diag((X'X))^{-1}})^{-1}(\mathbf{b} - \boldsymbol{\beta})$ are Student t variables.

Proof: From the perivious statment follos that $\sigma\sqrt{diag((X'X))^{-1}})^{-1}(\mathbf{b} - \boldsymbol{\beta})$ are standard normal variates. As

$$\begin{aligned}\mathbf{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{M}) &= \mathbf{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')) \\ &= \sigma^2((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')) = \mathbf{0},\end{aligned}$$

$\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}$ and $\mathbf{b} - \boldsymbol{\beta}$ are independent.

3.3.3 F-statistics

The distribution of the variance can be used to test several parameters together, essentially the relative quality of nested models.

F confidence regions If we have more than one parameter, and want to form a confidence region it is not clear what shape the region should have. Using F statistics naturally lead to ellipsoids.

As $\left[\frac{1}{\sigma^2}(\mathbf{X}'\mathbf{X})\right]^{\frac{1}{2}}(\mathbf{b}-\boldsymbol{\beta}) \sim \mathbf{N}(\mathbf{0},\mathbf{I})$, therefore

$$(\mathbf{b}-\boldsymbol{\beta})'\frac{1}{\sigma^2}(\mathbf{X}'\mathbf{X})(\mathbf{b}-\boldsymbol{\beta}) \sim \chi^2_k.$$

Dividing by $s^2/\sigma^2 = ESS/(n-k)\sigma^2$, we get

$$(\mathbf{b}-\boldsymbol{\beta})'\frac{1}{s^2}(\mathbf{X}'\mathbf{X})(\mathbf{b}-\boldsymbol{\beta})\frac{1}{k} = (\mathbf{b}-\boldsymbol{\beta})'\frac{1}{ESS}(\mathbf{X}'\mathbf{X})(\mathbf{b}-\boldsymbol{\beta})\frac{n-k}{k} \sim F_{k,n-k}.$$

This implicitly defines an ellipsoid for some positive number α.

$$(\mathbf{b}-\boldsymbol{\beta})'\frac{1}{ESS}(\mathbf{X}'\mathbf{X})(\mathbf{b}-\boldsymbol{\beta})\frac{n-k}{k} \leq \alpha.$$

F-tests Suppose our null hypothesis is that $\beta_2 = ...\beta_k = 0$ (only the constant is nonzero.) Then

$$\frac{RSS}{ESS}\frac{n-k}{k-1} = \frac{TSS-ESS}{ESS}\frac{n-k}{k-1} = \frac{R^2}{1-R^2}\frac{n-k}{k-1} \sim F_{k-1,n-k}.$$

According to the null TSS χ^2_{n-1} and we have seen that ESS is χ^2_{n-k}, moreover RSS and ESS are independent. Here $R^2 = 1 - \frac{ESS}{TSS}$.

By Cochrane's Theorem since

$$TSS = RSS + ESS,$$

RSS is χ^2_{n-1}.

It is not difficult to generalize for the null: $\beta_{j+1} = ...\beta_k = 0$. (Here we have $k-j$ restrictions.) Then

$$\frac{ESS_R - ESS_U}{ESS_U}\frac{n-k}{k-j} = \frac{R^2_U - R^2_R}{1-R^2_U}\frac{n-k}{k-j} \sim F_{k-j,n-k}.$$

If our hypotheses are not fomulated naturally as zero-restrictions we have another generalization.

Now the null can be formulated as

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r},$$

where $\mathbf{R}$ is an mxk $(m < k)$ matrix. .

Then

$$\begin{aligned} Var(\mathbf{Rb}) &= E(\mathbf{Rb})(\mathbf{Rb})') = \mathbf{R}\mathbf{E}(\mathbf{b}\mathbf{b}')\mathbf{R}' \\ &= \mathbf{R}\mathbf{V}ar(\mathbf{b})\mathbf{R}' \\ &= \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'. \end{aligned}$$

and $(\mathbf{Rb}-\mathbf{r})'(\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{Rb}-\mathbf{r}) \sim \chi^2_m$.
From which

$$\frac{(\mathbf{Rb}-\mathbf{r})'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{Rb}-\mathbf{r})}{ESS_U}\frac{n-k}{m} \sim F_{m,n-k}.$$

3.3.4 Properties of OLS with stochastic regressors

So far we have restricted our attention to the fixed $\mathbf{X}$ case. Clearly it is of interest to know how the above statements extend to stochastic regressors. Fortunately, quite well.

Unbiasedness is satisfied when the model estimates the CEF because $E(\epsilon \mid \mathbf{X}) = \mathbf{0}$ is fulfilled.

Consistency prevails quite generally (even for linear projection), though it requires certain assumptions on the stochastic properties of $\mathbf{X}$.

As the Gauss-Markov Theorem is valid for each $\mathbf{X}$, therefore it is also valid on average.

Strictly speaking $\mathbf{t}$ and $\mathbf{F}$ statistics are valid unconditionally only when there is joint normality. However, if we have a large enough sample they are correct asymptotically as the Central Limit Theorem implies that $\sqrt{n}(\mathbf{b}-\boldsymbol{\beta})$ converge to a zero mean normal vector:

$$\sqrt{n}(\mathbf{b}-\boldsymbol{\beta}) \sim asyN(\mathbf{0}, E(\mathbf{x}\mathbf{x}')^{-1}E(\mathbf{x}\mathbf{x}'\varepsilon^2)E(\mathbf{x}\mathbf{x}')^{-1}).$$

The standard errors are the square roots of the diagonal elements, and the covariance matrix simplifies in the case of homoskedasticity as $\sigma^2E(\mathbf{x}\mathbf{x}')^{-1}$.

3.3.5 Several possible regressions: what can the OLS estimate?

For the moment we only assume that

$$y_i = \boldsymbol{\beta}'\mathbf{x}_i + \epsilon_i,$$

and $E(\epsilon_i) = 0$.
We have four gears, in fact.
Gear 1: $E(\mathbf{x}_i\epsilon_i) = 0$.
In that case OLS estimates the parameters of the population regression (linear projection) consistently

$$p\lim_{n\to\infty}\mathbf{b}_n = \boldsymbol{\beta}.$$

Gear 2: $E(\epsilon_i \mid \mathbf{x}_i) = 0$. Then, in addition, OLS provides the unbiased estimates of the parameters of the linear CEF

$$E(\mathbf{b}) = \boldsymbol{\beta}.$$

Gear 3 (an extra):
Suppose

$$E(\epsilon_i^2 \mid x_i) = \sigma^2,$$

is also satisfied (homoskedasticity). Besides consistency and unbiasedness OLS is BLUE (minimum variance unbiased linear).

Gear 4: Suppose $(y_i, \mathbf{x}_i)$ are identical, jointly normal variables. Then OLS is globally efficient among unbiased estimators, and t, F tests can be conducted correctly.

3.3.6 The examples and OLS regression

The above four examples can be analyzed to show that each of them belongs to the different gears.

Example 1 The CEF and the population regression are different. The CEF residual is z, the regression residual $\epsilon = 3x^2 + z - 3$, thus $E(\epsilon \mid x) \neq 0$. Clearly the model is in Gear 1, OLS estimates consistently the linear projection, but not the CEF.

Example 2 Here both the CEF and the regression residuals are $3x^2-3$,and one can see that $E(\epsilon \mid z) = 0$. Therefore OLS estimates the CEF consistently and in an unbiased way (where the conditioning variable is z.) Also homoskedasticity is true, therefore the estimate is BLUE. As ϵ is not normal the model is only in Gear 3.

Example 3 The CEF and regression residuals are both z.The model is in Gear 4, as the joint distribution is normal.

Example 4 Here both residuals are xz. Clearly

$$E(\epsilon \mid x) = 0$$

therefore OLS estimates the CEF consistently and in an unbiased way, but it is not efficient, as homoskedasticity fails, since
$var(\epsilon \mid x) = x^2$.
Therefore the model is in Gear 2.

Heteroskedasticity (Gear 2) This is a very important case in practice. Here

$$\sigma_i^2 \neq \sigma^2.$$

The OLS estimator is unbiased and consistent, but not efficient.

Proposition: The OLS covariance matrix

$$\mathbf{var}(\mathbf{b}^{OLS}) = E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon})(\boldsymbol{\epsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1})).$$

is a biased and inconsistent estimator of the true covariance matrix.

This is the main problem with heteroskedasticity. Here the $E(\mathbf{X}'\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{X}) = \sigma^2 E(\mathbf{X}'\mathbf{X})$ simplification is not valid, and we need to estimate the $E(\mathbf{X}'\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{X})$ fourth-moment matrix.

Let us define S as

$$S = \frac{\sum_i \mathbf{x}_i \mathbf{x}_i^{'} u_i^2}{n}.$$

The heteroskedasticity-consistent covariance matrix estimator is:

$$var(\mathbf{b}^{OLS}) = (\mathbf{X}^{'}\mathbf{X})^{-1}\mathbf{S}(\mathbf{X}^{'}\mathbf{X})^{-1}.$$

The diagonal elements can be used for t tests.

Generalized Least Squares We look for a CEF estimate. The assumptions are now:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\mathbf{E}(\boldsymbol{\epsilon} \mid \mathbf{X}) = \mathbf{0}.$$

$$E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}' \mid \mathbf{X}) = \mathbf{diag}\left\langle \sigma_1^2, ..\sigma_i^2, ...\sigma_n^2 \right\rangle,$$

where σ_i^2 my depend on $\mathbf{X}$.

An even more general assumption is that

$$E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}' \mid \mathbf{X}) = \boldsymbol{\Omega}.$$

There exists a $\boldsymbol{\Omega}^{\frac{1}{2}}\boldsymbol{\Omega}^{\frac{1}{2}} = \boldsymbol{\Omega}$ decomposition for positive definite matrices. From this

$$(\boldsymbol{\Omega}^{-\frac{1}{2}}\epsilon)(\epsilon'\boldsymbol{\Omega}^{-\frac{1}{2}}) = \mathbf{I}.$$

Consider the transformed

$$\begin{array}{rcl} \boldsymbol{\Omega}^{-\frac{1}{2}}\mathbf{y} & = \boldsymbol{\Omega}^{-\frac{1}{2}}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Omega}^{-\frac{1}{2}}\boldsymbol{\epsilon} & \\ \mathbf{y}^* & = & \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^* \end{array}$$

model. This is homoskedastic and the OLS estimate

$$\mathbf{b}^{GLS} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y})$$

is unbiased, consistent, and efficient for $\boldsymbol{\beta}$, moreover

$$var(\mathbf{b}^{GLS}) = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}.$$

A simple subcase is the weighted least squares estimate when $\boldsymbol{\Omega}$ is diagonal.

Feasible GLS As Ω is unknown we need a consistent $\widehat{\Omega}$ estimate. One can have this from the OLS estimate. Having this we derive the feasible GLS estimator as

$$\mathbf{b}^{FGLS} = (\mathbf{X}'\widehat{\mathbf{\Omega}^{-1}}\mathbf{X})^{-1}(\mathbf{X}'\widehat{\mathbf{\Omega}^{-1}}\mathbf{y}).$$

3.4 Three general testing principles

An important theoretical concept is Fisher's information matrix:

$$\mathbf{F}(\mathbf{t}) = [-\mathbf{E}(\frac{\partial \log L}{\partial t_i \partial t_j})].$$

In the case of the normal regression model:

$$\ln L = -n \ln \sigma - n \ln(\sqrt{2\pi}) - \frac{1}{2\sigma^2} \sum (y_i - \mathbf{X}_i'\beta)^2,$$

and If $\mathbf{b} = \mathbf{b}_{ML}$. Then

$$\mathbf{F}(\mathbf{b}) = \mathbf{E}(\frac{X'X}{s^2}),$$

and

$$cov(\mathbf{b}_n) = \mathbf{F}(\mathbf{b})^{-1}.$$

In other words the inverse of the infomation matrix gives the covariance matrix of the ML estimator.

Let us consider the general restricted estimation problem:

$$\mathbf{R}(\boldsymbol{\beta}) = \mathbf{r},$$

where the Jacobi matrix of R is:

$$J(\mathbf{R}) = [\frac{\partial R_j}{\partial \beta_k}].$$

We have three general testing principles that are asymptotically equivalent.

3.4.1 The Likelihood Ratio principle

The distance between the restricted and the unrestricted estimate is measured as

$$\xi_{LR} = -2(\log L(b_U) - \log L(b_R)),$$

the log point difference between the likelihood values of the two estimates. (Approximately the percentage difference.)

3.4.2 Wald principle

The distance is:

$$\xi_W = (R(b_U) - r)' J(R_{b_u}) F_{b_U} J(R'_{b_u})(R(b_U) - r),$$

It is a distance of the estimated vectors in log points, where the distance is defined by the Jacobi.

3.4.3 LM principle

A new metric is introduced as

$$\xi_{LM} = \boldsymbol{\lambda}' J(R_R) F_{b_R}^{-1} J(R'_R) \boldsymbol{\lambda} = [\frac{\partial \log L}{\partial \mathbf{b}_R}]' F_{b_R}^{-1} [\frac{\partial \log L}{\partial \mathbf{b}_R}],$$

where

$$\frac{\partial \log L}{\partial \mathbf{b}_R} - J(\mathbf{R}'_R)\boldsymbol{\lambda} = 0.$$

This is a log point difference of the Lagrange-multpliers of the two estimates.

It can be proved that the three tests are asymptotically equivalent and distributed as χ^2_J.

A partial explanation of this theorem is that the Wald and LM test statistics are approximations to the LR statistic.

Let $L : R^n \to R$ be differentiable, and $\mathbf{\Delta L}_{x_0} = \mathbf{0}$.

$$L(x_1) - L(x_o) \cong \frac{1}{2}(x_1 - x_0)' H_{L_{x_o}} (x_1 - x_0).$$

where H is the Hessian. Therefore

$$2(L(x_1) - L(x_o)) \cong (x_1 - x_0)' H_{L_{x_o}} (x_1 - x_0).$$

This "explains" the asymptotic equivalence of LR and W, if L is the log-likelihood, x_0 is the unresricted ML estimator, and x_1 is the restricted ML estimator.

Moreover

$$\begin{aligned}
\mathbf{\Delta L}_{x_0} - \mathbf{\Delta L}_{x_1} &\cong H_{L_{x_1}}(x_0 - x_1) \\
-H_{L_{x_1}}^{-\frac{1}{2}} \mathbf{\Delta L}_{x_1} &\cong H_{L_{x_1}}^{\frac{1}{2}}(x_0 - x_1) \\
\mathbf{\Delta L}'_{x_1} H_{L_{x_1}}^{-1} \mathbf{\Delta L}_{x_1} &\cong (x_0 - x_1)' H_{L_{x_1}} (x_0 - x_1),
\end{aligned}$$

"explaining" that LM is asymptotically equivalent with the other two.

Notice that the usual Wald F test can be obtained from ξ_W by adjusting for the degrees of freedom.

ξ_{LM} can be computed from an auxiliary regression, where the target is the estimated residual from the restricted model, and the regressors include all

regressors in the general model. If R_a^2 is the coefficient of determination of the auxiliary regression, then

$$\xi_{LM} = nR_a^2.$$

In the case of multiple regression with linear restrictions:

$$\begin{aligned} \xi_{LR} &= n\log(\frac{ESS_R}{ESS_U}) \\ \xi_W &= n\frac{ESS_R - ESS_U}{ESS_U} \\ \xi_{LM} &= n\frac{ESS_R - ESS_U}{ESS_R}. \end{aligned}$$

$$\xi_{LM} \leq \xi_{LR} \leq \xi_W.$$

3.5 Literature

Green, W. H. (2003). Econometrics analysis (5e). Upper Saddle River, HNJ: Prentice Hall, 283-334.

Wooldridge, J. M. (2002). Econometric analysis of cross section and panel data MIT Press. Cambridge, MA, 108.

4 Structural estimation problems

Suppose: that

$$y = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3,$$

where y can be crop yield per area or earnings per month, x_1 hours with sunshine or years of education, x_2 water absorbed per area or the IQ of the worker, x_3 phosphate content of ground or stamina of the worker. Econometricians have always been concerned with the estimation of similar relationships, which were called structural equations. Probably the most traditional structural relationships economists have studied are the supply and demand functions. What makes a relationship "structural" is its character with respect to statistical assumptions.

A relationship is structural if it is valid irrespective of the "probability structure". In other words we can write down this equation without specifying anything about the random properties of the quantities involved. When we make assumptions about the distributions, too, then we transform this model into a statistical (probability) model. However, this transformation is not unique, and depending on it, we can obtain different results concerning the identifiability (estimability) of the parameters.

In the following let us assume that x_2 and x_3 are normal variates, x_3 is non-observed and has mean 0, while x_1 and x_2 can be observed. We are interested in estimating α_1. By setting the distribution of x_1 in different ways we obtain different models.

Case 0 (nature) x_1 is normal jointly with the other xs. Then

$$E(y \mid x_1, x_2) = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 E(x_3 \mid x_1, x_2),$$

and α_1 can be estimated by OLS from data consistently if and only if $E(x_3 \mid x_1) = 0$.

In this case we are exposed to the mercy of nature.

Case 1 (random experiment) We are able to set x_1 independently of anything relevant.

$$x_1 = u,$$

where u is independent of x_2 and x_3.Then the OLS estimate of α_1 is independent of the other variables, and it is consistent.

Case 2 (conditional independence assumption, see later the explanation) Here we are not able to set x_1 fully according to our wishes, and it is unavoidable that x_1 is correlated with the observable x_2, for instance

$$x_1 = \phi x_2 + u,$$

and $E(u \mid x_2) = 0$. However, if we are lucky and x_2 and x_3 are independent, then

$$E(y \mid x_1, x_2) = \alpha_1 x_1 + \alpha_2 x_2,$$

and α_1 is again recoverable from the data by OLS. But because of collinearity between x_1 and x_2 the estimator has a higher variance than in the former case.

Case 3: (selection bias) It is the unlucky case. However hard we try x_1 is not independent of the unobserved x_3.

$$x_1 = \xi x_3 + u,$$

and $E(u \mid x_2, x_1) = 0$
Then

$$E(y \mid x_1, x_2) = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 E(x_3 \mid x_1, x_2),$$

and

$$E(y \mid x_1, x_2) = (\alpha_1 + \frac{1}{\xi}\alpha_3) x_1 + \alpha_2 x_2.$$

The "true" coefficient α_1 is not recoverable from the data by OLS. In this example x_3 is called a confounder.

4.1 What is a causal effect?The potential outcome framework

Structural problems are essentially equivalent to causal estimation problems. The main (apparent) difference is that causal problems usually involve a causal variable that can take on a finite number of different treatment values. Causal problems are usually set in the potential outcome framework.

In a binary treatment case for the ith unit Y_{i0} is the potential outcome when $D_i = 0$ (no treatment), and Y_{i1} is the potential outcome when $D_i = 1$ (treatment)

The observed outcome is

$$Y_i = Y_{i0} + (Y_{i1} - Y_{i0}) D_i,$$

and the causal treatment effect can be defined as

$$(Y_{i1} - Y_{i0}) = \kappa.$$

It follows that

$$\begin{aligned} E(Y_{i1} \mid D_1) - E(Y_{i0} \mid D_0) = \\ E(Y_{i1} \mid D_1) - E(Y_{i0} \mid D_1) + \\ E(Y_{i0} \mid D_1) - E(Y_{i0} \mid D_0) \end{aligned}$$

In other words the average "observed" difference = average treatment effect + selection bias.

An example for selection bias is the case when patients with a better chance to recover get treatment in a medical experiment with higher probability than those with worse chances. We may attribute erroneously the better state of the treated patients to the effect of treatment. Our goal is to recover the average causal effect $E(Y_{i1} \mid D_1) - E(Y_{i0} \mid D_1)$ from the observations, by making the selection bias 0. One can guess that this case is formally equivalent to having a confounder in the structural problem.

In the following we always assume that the SUTVA (stable unit value) assumption is satisfied. It means that potential outcomes across individuals are independent. (One patient's state does not affect the state of any other, and there are no common influences that affect all patients.) This assumption is rather dubious in many economics applications (for instance if we want to estimate the effect of subsidies on firm performance.)

From now on we generalize to more than two treatment states. Suppose that

$$Y_i = C + \kappa D_i + \eta_i.$$

It can be regarded as a structural assumption without any reference to the distribution of D.

4.1.1 Random assignment

If D_i is independent of η_i then

$$\begin{aligned} E(Y_i \quad &| \quad D_i) = C + \kappa D_i. \\ \beta \quad &= \quad \kappa. \end{aligned}$$

Random assignment amounts to D_i being independent from anything that can affect Y_i. In that case OLS would recover κ consistently. In practice it is advisable to check whether the randomization is successful, which means that one must ask whether each level of treatment is represented uniformly in the sample (balance checking).

4.1.2 CIA (conditional independence assumption) and real human experiments

From now on treatment is characterizes by multiple values and $\mathbf{D}$ represents the corresponding vector of variables.

Frequently samples depend on some variable, for instance in schooling experiments participating schools are usually self-selected, but classes within schools can be chosen randomly. An assumption that can substitute for random assignment is as follows.

The conditional independence assumption (CIA): For any observable $\mathbf{X}$, relevant for the potential outcomes, potential outcomes and $\mathbf{D}$ are independent,

conditioned on **X**. This assumption corresponds to the case in the structural problem when the variable of interest was correlated only with relevant observed variables.

Then the

$$Y = \boldsymbol{\kappa}\mathbf{D} + \boldsymbol{\beta}\mathbf{X} + \eta_i$$

regression would give $\boldsymbol{\kappa}$ (a vector) as the causal effects. The CIA means that **X** is the only source of dependence between treatment assignment and the potential outcomes. It is important that the specified functional form must be correct.

We can classify explanatory variables in the following way:

a) The treatments **D**

b) Controls that are connected to the treatment. Importantly: all such variables must be present in the regression, these must constitute part of **X**.

c) Controls that are independent of the treatment, but are relevant. They belong also to **X**.

One can observe here a paradox: the estimate of the parameter of interest becomes more precise if relevant orthogonal variables are added to the regression. On the other hand if we add irrelevant variables the redundant variable problem arises.

An important point is that there might exist *bad control* variables. A variable that is influenced by the variable of interest, but does not affect the selection can be called a bad control, since if we include such a variable in the regression part of the total effect of the treatment will be attributed to it. We want to retain that pathway for the estimate of the causal effect. For instance in an educational experiment pre-experiment test scores can be included, but attrition rate cannot, if the response is the post-experiment test score.

Regression and causality: a practical guide I. Divide variables into observables and non-observables

Observables contain the

- outcome (y) (variable to be explained causally)
- treatment (D) (the potentially causal cariable)
- necessary control (X) (treatment assignment partly depends on it, and it also affects outcome)
- possible control (W) (independent of treatment, but may affect outcome)
- bad control (Z) (affected by treatment, can-be outcome).

Non-observables include

- confounders (C) (affect treatment assignment and outcome) and
- honest non-observables (u) (affect outcome, but independent of anything else)

II. One can formulate the following rules:

(1) If C is present linear regression does not work for recovering the causal effect.

(2) A truly random experiment excludes C and X.

(3) Otherwise X should be among the right-hand side variables.

(4) Z should not be among the right-hand side variables.

(5) Inclusion of W depends on judgement. It may increase imprecision if it does not affect y, but incrase precision if it does.

In any case the functional form must be approximately correct. But for linear regression linearity in parameters is what matters, thus it encompasses a wide range of non-linear functional forms in variables.

4.2 Matching: an alternative to regression

When applying the matching methodology the fundamental assumption is that we can identify (almost) identical individuals measured by relevant input characteristics (the X variables). The CIA is called in this literature the serendipity assumption: nothing essential is left out. Then the difference in the behaviour of a matched pair is truly random. There is another assumption needed: common support, which means that the probability of being treated or non-treated is non-zero for the same X. The basic case is full matching, when for each treated individual there exists at least one untreated with the same X properties.

The simplest estimate of the average causal effect on the treated is

$$\frac{1}{N_T}\sum_{i=1}^{N_T}(y_i(X_i, D_i = 1) - y_i(X_i, D_i = 0)).$$

Other estimators are possible, each of them takes some weighted average of the $y_i(X_i, D_i = 1) - y_i(X_i, D_i = 0)$ differences.

The main advantage of matching is that there is no need to find the correct functional form. The principal problem with matching is securing that the basic assumption is fulfilled. For this X must contain many variables, making less and less likely that exact matching can be achieved. Common support is also jeopardized if we increase the number of variables. In case of continuous $\mathbf{X}$s, it is practically impossible to satisfy.

There are several practical solutions for these problems. 1. One can apply approximate matching, based, for instance, on the Mahalanobis distance. 2. Approximate matching can be defined by the propensity score. This latter has a foundation in the following statement: if the CIA is satisfied with X, then it is satisfied with $p(X)$, where $p(X)$ is the probability of treatment conditioned on X (the propensity score). The propensity score must be estimated from data, where logit is the most frequently applied methodolgy.

In the simplest logit model the outcome (y) is binary (0 or 1). The fundamental assumption is

$$\begin{aligned} \log(\frac{P_i}{1-P_i}) &= \beta' x_i, \\ P_i &= \frac{\exp(\beta' x_i)}{1+\exp(\beta' x_i)}, \end{aligned}$$

where P_i is the probability of $y_i = 1$. The likelihiod function is the product of these probabilities assuming independence. Then

$$E(y_i \mid x_i) = P_i.$$

A possible algorithm for a matching strategy is the following:
(1) Choose X.
(2) Create matched samples with different methods.
(3) Check balance with each method.
(4) Prefer the matched data set with the best balance.
(5) Calculate the causal effect by some weighted average of the matched differences, and test significance.

4.3 Instrumental variables and causality

The origin of the instrumental variable estimation idea comes from the general problem of noisy observation of covariates.

4.3.1 Error in-variables problem

Suppose that

$$E(y \mid x) = \beta x,$$

but x_i can only be observed with noise:

$$x^* = x + u,$$

where u_i is the noise with properties

$$\begin{aligned} E(u \mid x) &= 0 \\ E(u^2) &= \sigma^2 \\ E(yu) &= 0. \end{aligned}$$

Consider the

$$y = \beta^{'} x^* + \epsilon$$

population regression. Then

$$\begin{aligned} \beta' &= \frac{E(x^* y)}{E(x^{*^2})} = \frac{E(xy)}{E(x^2) + \sigma^2} \\ abs(\beta') &= abs(\frac{E(xy)}{E(x^2) + \sigma^2}) < abs(\frac{E(xy)}{E(x^2)}) = abs(\beta). \end{aligned}$$

The OLS estimate b estimates consistently β':

$$b = \frac{\mathbf{x}^{*\prime}\mathbf{y}}{\mathbf{x}^{*\prime}\mathbf{x}^{*}} = \frac{\mathbf{x}^{*\prime}\mathbf{y}}{n} : \frac{\mathbf{x}^{*\prime}\mathbf{x}^{*}}{n}$$

$$\beta' = E(b) = p\lim b,$$

but is an inconsistent estimate of β, biased towards 0:

$$abs(p\lim b) < abs(\beta).$$

The problem is that β is the parameter of interest. What can we do? We can look for an observable z with the following properties:

$$\begin{aligned} E(zx) &\neq 0 \\ E(zu) &= 0 \\ E(z\epsilon) &= 0. \end{aligned}$$

Then:

$$\begin{aligned} E(zy) &= \beta E(zx^{*}) \\ \beta &= \frac{E(zy)}{E(zx^{*})}. \end{aligned}$$

Estimating the parameters with sample moments we get

$$\begin{aligned} b_z &= \frac{\mathbf{z}'\mathbf{y}}{\mathbf{z}'\mathbf{x}}. \\ p\lim b_z &= \beta. \end{aligned}$$

4.3.2 Structural (causal) estimation with instrumental variables

The IV idea Suppose the CIA is not satisfied, i.e. there is at least one confounder. A possible instrument is any variable that has no role in the supposed structural relationship, is independent of the confounder, but is correlated with the variable of interest. More formally the setup is the following:

$$y = \beta x + \eta$$

and

$$\begin{aligned} \eta &= cw + u \\ cov(x, w) &\neq 0 \end{aligned}$$

Then

$$cov(x\eta) \neq 0.$$

In other words w is a confounder, if we want to estimate β from a sample that do not contain observations on w.

If there exists z (called an instrument for x) which satisfies

$$cov(z, x) \neq 0$$

(relevance)

$$cov(z, \eta) = 0$$

(uncorrelatedness)
and

$$\begin{aligned} E(y \quad | \quad x, z) &= \beta' x + \gamma' z \\ \gamma' \quad = \quad 0. \end{aligned}$$

(exclusion)
then

$$cov(yz) = \beta cov(xz),$$

and therefore

$$\beta = \frac{cov(yz)}{cov(xz)},$$

where β is the looked for causal effect of x on y.

For example a typical labour economics problem is the following. For each individual let y be earnings, x the length of education, w abilities, and z the month of birth. The variable of interest is the length of education but it is related to abilities, an unobserved variable, which also affects earnings in its own right. The relevance of month of birth is satisfied if length of education depends on the month of birth, which can be proven sometimes empirically. Independence is satisfied plausibly as month of birth and abilities are thought to be independent. The exclusion restriction is satisfied if the only effect of month of birth on earnings is via the length of education, which is a plausible assumption.

This problem can also be formulated in the traditional simultaneous structural equations framework in econometrics. In this somewhat special case the "structural" form consists of two equations:

(1) the population regression of x on the instrument:

$$x = \gamma z + u,$$

and (2) the structural (causal) equation:

$$y = \beta x + \varepsilon',$$

where $\varepsilon' = \gamma\eta + \varepsilon$, and which is not the conditional expectation function as $cov(x\varepsilon') \neq 0$, since $cov(x, \eta) \neq 0$.

Then we obtain the reduced form by solving the structural equations in terms of z:

$$x = \gamma z + u$$

$$y = \beta\gamma z + (\varepsilon' + u) = \delta z + u'$$

where $cov(z, u) = 0, cov(z, u') = 0$. Thus both equations are population regressions on z.

From these:

$$\beta = \frac{\delta}{\gamma},$$

provided that $\gamma \neq 0$ (the coefficient of z is non-zero in the first equation). This is another route to estimate the causal effect (called indirect LS).

A third way to achieve exactly the same outcome exists, too. Define the projected value as a random variable

$$\widehat{x} = \gamma z.$$

Then $\widehat{x}$ is also a valid instrument by definition and

$$\begin{aligned} cov(\widehat{x}, y) &= \beta cov(\widehat{x}, x) \\ \beta &= \frac{cov(\widehat{x}, y)}{cov(\widehat{x}, x)}. \end{aligned}$$

Or alternatively:

$$\beta = \frac{cov(\widehat{x}, y)}{var(\widehat{x})},$$

since $var(\widehat{x}) = var(x)$.

The first formula shows that β is the parameter on $\widehat{x}$ in a population regression where we regress y on $\widehat{x}$. Therefore it is called two-stage LS (2SLS). In the first stage we create $\widehat{x}$, then with $\widehat{x}$ we do another regression, and the wanted parameter is the parameter on $\widehat{x}$ in the second-stage regression.

$$\beta = \frac{cov(\widehat{x}, y)}{var(\widehat{x})} = \frac{cov(y, z)}{cov(x, z)} = \frac{cov(y, z)}{var(z)} : \frac{cov(x, z)}{var(z)}.$$

2SLS has the additional attraction that it can be generalized for several instruments. Suppose

$$x = \gamma_1 z_1 + \gamma_2 z_2 + u,$$

where z_1 and z_2 are valid instruments, is a population regression.

The reduced form in this case consist of:

$$x = \gamma_1 z_1 + \gamma_2 z_2 + u$$

and

$$y = \beta(\gamma_1 z_1 + \gamma_2 z_2) + (\beta u + \varepsilon') = \delta_1 z_1 + \delta_2 z_2 + \epsilon'.$$

Then

$$\widehat{x} = \gamma_1 z_1 + \gamma_2 z_2$$

is also a valid instrument, and

$$\beta = \frac{cov(y, \widehat{x})}{var(\widehat{x})}.$$

This is called the overidentified case of 2SLS.

If both $\widehat{x}_1$ and $\widehat{x}_2$ (the instruments created from one-variable population regressions) are valid then $\widehat{x}$ is more efficient.

Structural linear regression in the general case with mathematical formulas Suppose that

$$y = \boldsymbol{\beta}\mathbf{x} + \varepsilon$$

where $E(\mathbf{x}\epsilon) \neq 0$. and $\boldsymbol{\beta}$ is the parameter of interest.

Then

$$\boldsymbol{\beta} \neq \mathbf{E}(\mathbf{x}\mathbf{x}')^{-1}\mathbf{E}(\mathbf{y}\mathbf{x}').$$

The IV estimate If there exist $\mathbf{z}$ with the same dimension as $\mathbf{x}$, $\mathbf{E}(\mathbf{z}\boldsymbol{\epsilon}) = \mathbf{0}$,and $\mathbf{E}(\mathbf{x}\mathbf{z}')$ non-singular then

$$\mathbf{E}(\mathbf{y}\mathbf{z}') = \boldsymbol{\beta}\mathbf{E}(\mathbf{x}\mathbf{z}'),$$

and therefore

$$\boldsymbol{\beta} = \mathbf{E}(\mathbf{x}\mathbf{z}')^{-1}\mathbf{E}(\mathbf{y}\mathbf{z}').$$

This is the population relationship whose sample equivalent is:

$$\mathbf{b}^{iv} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$$

and plim$\mathbf{b}^{iv} = \boldsymbol{\beta}$, if $plim\frac{\mathbf{Z}'\mathbf{X}}{n}$ non-singular, $plim\frac{\mathbf{Z}'Z}{n}$ positive definite, and $plim\frac{\mathbf{Z}'\varepsilon}{n} = 0$.

We can divide $\mathbf{x}$ into two parts:

$$\mathbf{x} = [x_1, x_2],$$

where

$$\begin{aligned} E(\mathbf{x}_1\epsilon) &\neq 0 \\ E(\mathbf{x}_2\epsilon) &= 0. \end{aligned}$$

Then the $\mathbf{x}_1$ variables are called endogenous, while the $\mathbf{x}_2$ variables exogenous. The $\mathbf{x}_2$ variables are their own instruments.

Indirect LS Consider

$$\mathbf{B}_{x,z} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$$

and

$$\mathbf{b}_{y,z} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}.$$

Obviously

$$\mathbf{b}^{iv} = \mathbf{B}_{x,z}^{-1}\mathbf{b}_{y,z}.$$

2SLS Now z's dimension is at least as large as x's. Consider the regression of x on z.

$$\widehat{X} = \mathbf{B}_{x,z}\mathbf{Z}.$$

Then $\widehat{x}$ is another possible set of instruments and

$$\mathbf{b}^{2sls} = (\widehat{X}'\mathbf{X})^{-1}\widehat{X}'\mathbf{y}$$

consistent.
As

$$\begin{aligned} (\widehat{\mathbf{X}}'\widehat{\mathbf{X}}) &= \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} = \\ \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} &= (\widehat{\mathbf{X}}'\mathbf{X}) \end{aligned}$$

the estimator can be written as

$$\mathbf{b}^{2sls} = (\widehat{X}'\widehat{X})^{-1}\widehat{X}'\mathbf{y}.$$

In other words $\mathbf{b}^{2sls}$ is the OLS parameter vector from regression of $\mathbf{y}$ on $\widehat{X}$.

How to calculate standard errors with 2SLS estimation? Standard errors are not to be calculated from second-stage residuals, but from

$$RSS_{IV} = (y - X\mathbf{b}^{2sls})'(y - \mathbf{X}\mathbf{b}^{2sls})$$

the "true" 2SLS residuals ($X\mathbf{b}^{2sls} \neq \widehat{\mathbf{X}}\mathbf{b}^{2sls}$.) Then the $t(z), \chi^2$ and F tests are asymptotically valid.

Diagnostic testing 1. Relevance (are instruments correlated with the causal variables?) can be tested with an F test from the reduced form. It is also called a weak instrument test.

2. One can test whether the causal variables correlate with the structural residuals (endogeneity test). If they do not then the IV is meaningless, and one should simply estimate the structural equations with OLS. For the test the estimated residuals of the reduced form are put into the structural equation as explanatory variables. If the parameters are not significantly different from 0 then the whole IV procedure is futile (the Wu-Hausman test).

3. Do all the instruments satisfy the exclusion conditions? (Overidentification or Sargan test.) If not all of them are valid instruments then some of them must enter the equation as its own instrument. For the test 2SLS residuals are regressed on all explanatory variables and instruments. The null is that each coefficient in this regression is 0. If the null is rejected one must reduce the number of instruments, but it is not obvious exactly how.

A practical guide to IV modelling of structural effects 1. Select a response y, endogenous variables $\mathbf{x}_1$(whose effects are the parameters of interest), $\mathbf{x}_2$ exogenous controls, and write down the structural equation.

2. Choose instruments $\mathbf{z}$ for the endogenous variables. The number of instruments must be larger than the number of endogenous variables.

3. Estimate the structural parameters by 2SLS. Test parameter significance.

4. Conduct the test for weak instruments. If the null is accepted look for other instruments.

5. Test the endogeneity of endogenous variables. If the null is accepted you can reestimate with OLS.

6. Test overidentification restrictions. If the null is rejected try to find out which instruments can be changed to controls.

4.4 Regression discontinuity design (RDD)

This is a methodology that identifies the treatment effect via the discontinuity of treatment probability as a function of some variable. The idea is that on the two sides of the discontinuity individuals are "almost the same", at least close to the point of discontinuity.

4.4.1 Sharp RDD

The legal drinking age constitutes a chilling discontinuity in the probability of death for American youth. In this case treatment (having the right to drink alcohol) is a deterministic function of the running (forcing) variable, x, which is age.

$$\begin{aligned} D &= 0, x \leq x_0 \\ D &= 1, x > x_0 \end{aligned}$$

where $D = 1$ is the legal right to buy alcoholic liquors, and x_0 is the legal age.

4.4.2 Parametric sharp RDD

If we are confident we can define a regression:

$$y = \beta D + P(x) + \epsilon,$$

where $P(x)$ is a polynomial in x. Or, more generally,

$$y = \beta D + P_0(x) + \gamma D P_1(x) + \epsilon.$$

Here β and γ are the parameters of interest. Implicitly the CIA is assumed here: the only variable that affects treatment is x. The functional form of the polynomial must be correct, however, that's why confidence is needed.

4.4.3 Non-parametric sharp RDD

If we are less confident, and we do not trust in our ability to get the functional form right we can define windows around the cutpoint x_0, and take local averages above and below. The difference of the averages is the the treatment effect approximately, and it is independent of the functional form. "Averaging" is not an unique concept, different estimators use different weighting schemes.

4.4.4 Fuzzy RDD

In some cases the running variable does not determine fully the treatment, but there is a sharp change in treatment probability at the cutoff point. Here the parametric structural assumption looks like:

$$y = \beta D + \gamma X + \epsilon,$$

where X contain other exogenous variables (possibly polynomials in x), but it can happen also that D is endogenous (i.e correlated with ϵ).

We assume that

$$\begin{aligned} P(D &= 0) = g_0(x), x \leq x_0 \\ P(D &= 1) = g_1(x), x > x_0 \\ g_0(x) &\neq g_1(x), \end{aligned}$$

in other words the running variable x does not determine D deterministically, while the probabilities of being treated are affected by the running variable, and there is a discontinuity at x_0. We guess that there is some confounder that influences D and y as well. However, we have as instrument T:

$$\begin{aligned} T &= 0, x \leq x_0 \\ T &= 1, x > x_0, \end{aligned}$$

the assignment for the treatment. As some people will decide against treatment, and others might get treatment without being assigned assignment and treatment do not coincide. Still assignment depends only on the observable x, and T is a valid instrument since it is related to D, unrelated to any confounder, and, in itself, does not affect $\mathbf{y}$. This is therefore an IV version of regression discontinuity, and must be dealt with accordingly. The first stage determines treatment as a function of the instrument, and, possibly, of other exogenous variables.

$$D = \lambda + \theta T + \xi X + u,$$

The second-stage is:

$$y = \alpha + \beta \widehat{D} + \gamma X + \epsilon.$$

A non-experimental, mathematically equivalent, example is an educational problem. Suppose y is seventh-grade test score, D seventh-grade peer test score, and R is the individual's fourth-grade test score. The causal problem is whether peer quality is important for the performance of students.

If we ran a naive regression

$$y = \alpha + \beta D + \gamma R + \epsilon,$$

the parameter of interest β would appear to be significant, but D is probably "endogenous", i.e. there must exist unobserved variables that affect both individual and peer seventh-grade performance.

We should look for a good instrument. Q qualification (in the past) for a "good" school (a dummy) is a valid instrument of D. Qualification meant having an entry exam score above some threshold value, it is clearly correlated with R, and is very likely correlated with D. On the other hand, being in the past, supposedly does not have an independent influence on the current test score, and on nothing else that may affect it.

The complete model consists of the equations:

$$\begin{aligned} D &= \delta Q + \xi R + u \\ \widehat{D} &= \widehat{\delta} Q + \widehat{\xi} R \\ Y_i &= \beta \widehat{D} + \gamma R + \epsilon. \end{aligned}$$

The parameter of interest is β, which may turn out to be insignificant.

A short guide to RDD analysis 1. Define an RDD object. What are the response, the running variable, the cutpoint, and other exogenous variables? If it is fuzzy then you must provide the instrument(s) as well.

2. Check the assumptions. For instance the running variable must be smooth around the cutpoint, the discontinuity must have its origin in treatment assignment alone.

3. Estimate the model parametrically or non-parametrically.

4. Check the sensitivity of results. It is especially important in the case of non-parametric estimation, where the width of the window is not an obvious choice.

4.5 Difference-in-Differences

This method of estimating causal effects works when we have data observed at different points of time, and we are able to eliminate time invariant confounders in some way. In this case we can observe entities in an untreated state at time T, and, while some of them stay untreated at T+K, some gets treatment later. When we take differences over both groups and then take the difference of the differences we obtain the causal effect, provided that some other conditions are also fulfilled.

4.5.1 Panel fixed effects models

We have panel data, in other words the same units can be observed over two or more time periods.

Our basic assumption can be formulated as:

$$E(y_{it} \mid A_i, t, X_{it}, D_{it}) = A_i + f(t) + \lambda X_{it} + \beta D_{it}.$$

Here $f(t)$ is a time trend common to all individuals (the common trend assumption), X_{it}s are individual specific observed exogenous variables, and A_i is a non-observed individual characteristic, which is therefore a confounder, as it may correlate with D_{it}(the treatment variable). This is called the fixed-effect panel model. Without the A_i there would be no problem with causal estimation, but in practice with non-experimental data confounding is always present. It is important to notice that here fixed effects (the confounders) are individual specific, but time invariant, and time effects $f(t)$ are common among individuals (the common-trend assumption).

We have two main ways to eliminate the confounding variables.and then estimate β.

1. Taking time averages over individuals. Then one can write the equation in deviation from averages. It "kills" the fixed effect, because the deviation from average is zero in the case of A_i. (It is called the *within* estimator).

2. Taking differences between periods. This again kills the fixed effects as

$$\nabla y_{it} = f(t+1) - f(t) + \lambda \nabla X_{it} + \beta \nabla D_{it} + \epsilon_{t+1} - \epsilon_t.$$

Obviously ∇D_{it} cannot be identically 0, there must exist changes in treatment status.

These two methods lead to different residuals. In the latter equation it is clear that there must be residual autocorrelation. With panel data homoskedasticity is usually not satisfied, estimation and, especially, tests must take this into account.

4.5.2 Groups and difference-in-differences (DID)

A leading example is the effect of minimum wages on employment in Philadelphia and New Jersey, when New Jersey introduced a new minimum wage in 1993. Researchers looked for changes in employment in fast-food restaurants in the two states to establish whether the minimum wage increase affected employment.

In this type of models individuals belong to groups (indexed by s). Confounding effects are present at the group level (a_s).

$$E(y_{ist} \mid s, t, D_{st}) = a_s + f(t) + \lambda D_{st}.$$

Then

$$E(y_{ist} \mid s, t{+}k, D_{s,t+k}) - E(y_{ist} \mid s, t, D_{s,t}) = f(t{+}k) - f(t) + \lambda \nabla D_{st} = diff(s, t, t{+}k).$$

Consider $s \neq s'$.

$$diff(s, t, t+k) - diff(s', t, t+k) = \lambda(\nabla D_{st} - \nabla D_{s't}).$$

If, for example, $\nabla D_{st} = 1$ and $\nabla D_{s't} = 0$, then λ can be estimated as

$$(av(Y_{is,t+k}) - av(Y_{is,t})) - (av(Y_{is',t+k}) - av(Y_{is',t}))$$

It is a "weighted regression" where groups' data are weighted by their relative size.

4.5.3 Regression DID

Suppose we have two groups (here fast-food restaurants in New Jersey and Philadelphia, respectively), and we define a group dummy D_s which takes the value 1 for the treated group. Then the previous model can be written as a regression:

$$Y_{ist} = \alpha + \beta D_s + \gamma D_t + \lambda D_{st} + \epsilon_{ist}.$$

Here D_t is a time dummy, with 0 for pre-treatment periods, and 1 for post-treatment periods, and

$$D_{st} = D_t D_s.$$

The parameter of interest is λ, measuring the effect of treatment on the treated.

This equation can be generalized by including exogenous variables X_{ist}, for several groups and periods. Indeed it is just a regression framework.

In sum, we can say that DID is applicable when

1. treatment has a time reference, and there are observations both pre- and after-treatment

2. confounders (relevant non-observed variables) can be "differenced-out" either at the group or at the individual level.

3. the common trends assumption can be maintained.

In addition:

4. There might still exist confounders either at the s or the i levels. Here we do not assume the CIA, therefore we must find instruments for the treatment variable. (Find a variable that is correlated with treatment but not correlated with the confounder, and has no place in the structural equation.)

4.6 Literature

Wooldridge, J. M. (2002). Econometric analysis of cross section and panel data MIT Press. Cambridge, MA, 108.

Angrist, J. D., & Pischke, J. S. (2008). Mostly harmless econometrics: An empiricist's companion. Princeton university press.

5 The inductive approach: statistical learning

5.1 Prediction

In many cases we want to predict the target value of an observation that does not belong to the sample from which we have calculated our estimates. Let $P(X)$ is an estimator, and let $\widetilde{y}_0$ be interpreted as the predicted value of the unknown y_0, $\widetilde{y}_0 = P(X_0)$.

To evaluate predictions we should make assumptions. Suppose that y_0 and X_0 are independent of the sample (X), and the conditional expectation function is $E(y_0 \mid X_0) = F(x_0)$. The mean squared error (MSE) is

$$\begin{aligned} MSE(x_0) &= E(y_0 - P(X_0))^2 = E(F(x_0) + \epsilon - P(X_0))^2 = \\ &= E(\epsilon^2) + E(E(P(x_0) - F(x_0))^2 + E(E(P(X_0)) - P(X_0))^2. \end{aligned}$$

The MSE is the sum of three terms: 1. The irreducible uncertainty, which is the consquence of x_0 being random. 2. The squared bias. This depends on the "quality" of the estimator in terms of unbiasedness. 3. The estimator's variance (following from the fact that the estimator is a random variable since the sample is random). This latter can be reduced by increasing the sample size. There might be a trade-off between the second and third terms. An unsophisticated model may be biased but may have little variance, whereas a sophisticated model may be unbiased but may have a large variance. In other words if we want to have a good prediction (in the sense of a small MSE) it is not necessarily the case that looking for an unbiased estimate of the CEF is the best idea.

Prediction is inevitably a problem of generalization. We want to have a statistical model that works well outside the sample, which is called training sample in this literature. On the other hand prediction must have a definite purpose. The statistical learning literature is based on the idea that good generalizations can be obtained by (learning) algorithms rather than setting up fixed assumptions about a problem, and proceeding by deduction from these. Traditional statistical practice does something similar implicitly, when diagnostic testing is applied and models are reformulated as a result of tests. The statistical learning literature carries out this program more systematically, and uses somewhat different concepts than the traditional literature. Loss functions, hyperparameters, training, validation and test samples are concepts that are employed incidentally by traditional statisticians, but here these are basic concepts. Also testing the generalization capabilities of a model is a must here, in contrast to the traditional t or F tests.

5.2 The problem setting

X is the input (feature, covariate, explanatory or exogenous variables) space. There exists a function $F : X \to Y$, where $y \in Y$ is the the target variable, the "true" relationship. The true relationship can be observed only with some noise, however. The true relationship can be taken as the expectation of Y

conditional on X. We want to get an estimator $P : X \to Y$ based on a finite data set that is "optimal" in the sense of giving optimal predictions for y based on information about $\mathbf{X}$. Determining optimality clearly requires an objective function which is usually a loss function in this literature.

Typical loss functions include the following.

Squared error:

$$L(Y, P(X)) = (Y - P(X))^2.$$

This is implicitly the loss function applied by traditional regression analysis, it is perfectly well suited to normal distributions.

Absolute error:

$$abs(Y - P(X)).$$

This criterion leads to the estimation of some median, rather than a mean. It is the only meaningful criterion with ordinal target variables.

Likelihood loss:

$$-2 \log \Pr(Y \mid P(X)).$$

This can be applied for all sort of targets, including qualitative variables. An obvious disadvantage is that the distribution must be known explicitly. For instance if the true model is Gaussian then

$$\Pr(\mathbf{y}, \mathbf{X}; \beta) = \sigma^{-N}(2\pi)^{-\frac{N}{2}} \exp(-\frac{\sum_{i=1}^{N}(y_i - \mathbf{X}_i'\boldsymbol{\beta})^2}{2\sigma^2}),$$

(Here Pr denotes the density.)

For classification problems, where Y_i qualitative ($G = 1, ..k....K$) loss functions include $0-1$ loss, or the likelihood loss which is called here deviance. Typically one sets up a model to explain $p_k(X)$ (the probability that type k is realized when X). Then the investigator assigns to X the type with the highest probability.

$$\widehat{G}(X) = \arg\max_k \widehat{p}_k(X),$$

where $\widehat{p}_k(X)$ is the estimate of $p_k(X)$.

With $0-1$ loss the loss is 1 when the classification is wrong, otherwise it is 0.

$$L(G, \widehat{G}(X)) = I(G \neq \widehat{G}(X)).$$

With likelihood loss:

$$L(G, \widehat{p}_G(X)) = -2 \log \widehat{p}_G(X).$$

In principle these cases are different.

5.2.1 Types of errors

We can distinguish among a number of different errors. The potentially most useful to measure is the generalization error:

$$Errg_T = E(L(y_0, P(X_0) \mid T)).$$

This is the expected loss of applying the prediction, estimated on a training sample (T), on a test observation which is independent of the training set. That is the ideal goal of estimation, but in practice this is normally unreachable.

The expected generalization error

$$Errg(x_0) = EL(y_0, P(X_0)).$$

is the expectation of the former, where the expectation is taken over all training sets. This can be estimated sometimes.

The ideal solution would beif we had a triple division of data: Training set - Validation set - Test set. On the training set we would estimate models with different tuning (complexity) parameters (also called hyperparameters). On the validation set we would choose the tuning parameter with the smallest expected generalization error. And finally on the test set we would estimate the generalization error for the model selected.

5.2.2 Information criteria: a surrogate for the generalization error

Suppose, however, that we do not have enough data to apply the ideal procedure. Though we would like to tell something about generalization errors as an intermediate step we must consider errors for each x_i in the training sample, too:

$$Err_T(x_i) = E(L(y_i, P(x_i) \mid T)).$$

It is a theoretical quantity, since it is an expected value.

We define the in-sample prediction error as

$$Err_{in} = \frac{1}{n} \sum_{i}^{n} Err_T(x_i),$$

which is also a theoretical quantity.

Then the training error is defined as the observed counterpart of the in-sample prediction error.

$$\overline{err} = \frac{1}{n} \sum_{i=1}^{n} L(y_i, P(x_i)).$$

One can guess that the training error (the average residual sum of squares in the case of a quadratic loss function) is an optimistic estimate of the in-sample error.

The bias-variance decomposition and linear regression The model in this case can be written as

$$\begin{aligned} y &= F(x) + \epsilon \\ E(\epsilon) &= 0 \\ var(\epsilon) &= \sigma^2. \end{aligned}$$

Suppose we estimate by OLS:

$$b^{OLS} = (X'X)^{-1}X'y.$$

The natural prediction function is:

$$\widetilde{y}_0 = P(X_0) = x_0' b^{OLS}.$$

For least squares linear regression the expected squared loss is:

$$err(\mathbf{x}_0) = \sigma^2 + (\mathbf{x}_0' E(b^{OLS}) - F(\mathbf{x}_0))^2 + (\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)\sigma^2.$$

If it is well specified (it is a CEF), then the bias is 0, but it is not necessarily the case.

The in-sample expected error is

$$Err_{in} = \frac{1}{N}\sum Err(x_i) = \sigma^2 + \frac{1}{N}\sum(\mathbf{x}_i'(E(b^{OLS})) - F(x_i))^2 + \frac{p}{N}\sigma^2.$$

Here in the formula the observations take the place of x_0. From least squares theory it is known that

$$\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

is idempotent and have rank p, therefore its trace equals p, the number of estimated parameters.

Model complexity, as indicated by p, is related to the in-sample error!

Optimism of the estimator can be defined as

$$op = Err_{in} - \overline{err}.$$

It can be proved that expected optimism for several loss functions satisfies

$$E(op) = \frac{2}{N}\sum cov(y_i, P(x_i)).$$

(The observed y_is are correlated positively with $P(x_i)$ (the estimator), whereas "new" observations at the same x_i are not.)

For the linear regression model

$$\sum_{i=1}^{N} cov(y_i, x_i' b^{OLS}) = trace(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\sigma^2 = p\sigma^2.$$

Thus expected optimism increases with the number of parameters, and with σ^2.

Definition of the effective number of parameters The definition is motivated by the previous formula:

$$df = \frac{\sum cov(y_i, P(x_i))}{\sigma^2}.$$

Therefore in linear regression:

$$df = p.$$

Estimating the in-sample prediction error

The C_R statistics We look for an estimate of the in-sample prediction error (a theoretical quantity):

$$\widehat{Err_{in}} = \overline{err} + \widehat{E(op)}.$$

Based on the above argument under squared error loss C_R is a reasonable estimate:

$$C_R = \overline{err} + 2\frac{p}{N}\widehat{\sigma^2}.$$

Akaike Information Criterion Suppose we consider likelihood loss. It can be proved that asymptotically:

$$-2E(\log \Pr_{\beta^{ML}} y) \simeq -\frac{2}{N}E(\sum \log \Pr_{\beta^{ML}}(y_i)) + 2\frac{p}{N}.$$

Let *loglik* be the maximized log-likelihood

$$\sum \log \Pr_{\beta^{ML}}(y_i) = \log lik.$$

Then the Akaike-criterion is

$$AIC = -\frac{2}{N}\log lik + \frac{2p}{N}.$$

For the Gaussian model likelihood loss and squared loss are the same, thus

$$C_R = AIC.$$

The Bayes factor and minimum description length Another estimate is the bayesian information criterion:

$$BIC = -2 \log lik + \log(n)p.$$

Obviously it is also designed for likelihood loss. Comparison of models via the BIC is equivalent to using the Bayes factor with uniform priors.

An alternative interpretation of the BIC is the following. When transmitting (optimally) a random variable about $-\sum \Pr(y_i) \log_2 \Pr(y_i)$ bit information is needed on average. The BIC compares two models (i.e. two ways of transmitting information on Y) via their minimum description lengths. Originally AIC, too, was derived from information theoretical principles.

To sum it up: if we must rely only on the training set we can still estimate the in-sample prediction error, using its observed equivalent the training error but correcting for its "optimism", and choose the best model based on it.

5.2.3 Validation: one step closer to the generalization error

We still do not think we can set aside data for testing, and calculating the generalization error. Cross-validation can be used for estimating the expected generalization error, and choosing the best model with the minimum expected validation error from a set of candidates. Another method of circumventing the data problem is to use bootstrap sampling.

K-fold cross-validation Let $\kappa : \{1, ...N\} \to \{1, , , K\}$ an indicator function that divides the data into K equal-sized subsets.

$f_\beta^{-k}(x, \lambda)$ is the function indexed by the tuning parameter λ, where estimation uses data not belonging to the kth fold. Then we can calculate the cross-validation statistics for each λ :

$$CV(f_\beta, \lambda) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f_\beta^{-\kappa(i)}(x_i, \lambda)).$$

The chosen model is with the λ that minimizes $CV(f_\beta, \lambda)$. The typically choices are $K = 5$ or 10. If $K = N$, it is called leave-one-out cross validation. The choice of K is empirical and casual. There is an obvious trade-off: if K is large the computational burden is substantial and there is large variance (as training sets are similar), if K is small there are only few estimates to average over (the Law of Large numbers does not have enough force.)

The bootstrap as validation Bootstrap sampling means that from a sample of size N we create many (T) new random samples, each of size N, by random selection and replacement. For each λ we calculate

$$CV(f_\beta, \lambda, boot) = \frac{1}{T} \sum_{j=1}^{T} (\frac{1}{N} \sum_{i=1}^{N} L(y_i, f_\beta^j(x_i, \lambda))),$$

where j indexes the bootstrap samples. We choose the model with the λ that minimizes $CV(f_\beta, \lambda, boot)$.

5.3 Machine learning algorithms

5.3.1 Regression learning algorithms

A typical way to find a good predictor is to start with a set of functions parametrized by β (for instance linear regression with β parameter vector), and a loss function $L(Y, P_\beta(X))$, where P_β is an element of this set. It may be reasonable to minimize a perturbed loss function $E_X L(Y, P_\beta(X)) + h(\beta, \lambda)$, where λ (a hyperparameter) measures the complexity of the function parameterized by β, since, as we have seen, more complex functions (with more effective parameters) have larger variance.

Ridge regression Ridge regression is a a generalization of OLS. We wish to minimize the following criterion function:

$$SSR + \lambda \boldsymbol{\beta}' \boldsymbol{\beta}.$$

where SSR is the sum of squared residuals, and λ is a complexity parameter. (λ penalizes large $\boldsymbol{\beta}$, if it deviates from 0 too much.) Plausibly the result is that the estimate is shrunken towards 0:

$$\mathbf{b}^{ridge} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}'\mathbf{y}.$$

(As the inverse of a matrix decreases in absolute value if any element of it increases in absolute value.) This method wants to exploit the bias-variance trade-off, the optimal λ may be set by cross-validation.

The Lasso The Lasso differs from ridge regression in the perturbed criterion function:

$$SSR + \lambda \sum_{k=1}^{K} \mid \beta_k \mid .$$

The penalty is formulated as a sum of absolute deviations from 0, rather than squared deviations. A disadvantage of Lasso is that explicit solution does not exist. However the Lasso makes variable selection for large enough λ, as certain parameters may become 0, which is generally considered as an advantage. The selection of λ occurs via cross-validation, again.

5.3.2 Tree-based methods

Growing a tree recursively To be concrete the detailed description below addresses the binary classification problem with negative entropy as a measure of the goodness of fit, at the end we give the necessary modifications for other types of models.

Output data define a binary distribution over the two classes they belong to. This distribution has an entropy, reflecting the uncertainty one faces when wishing to classify the objects without the knowledge of any explanatory variables. Tree growing is in essence an entropy reduction process. At the beginning consider each explanatory variables (features) and calculate by how much total entropy would be reduced if one were to split the full sample in two, based on the variable in question. If a variable has many possible values then there are many splits to consder, and one must choose the one with the highest reduction in entropy. After considering each variable in turn, select the one with the highest entropy reduction capacity, and perform the corresponding split of the sample. Graphically this is equivalent to forming two nodes in a tree whose parent node is the root. Geometrically a partition of the input space is the result. Entropy reduction can be viewed alternatively as purifying: the new nodes are purer than the root node, in the sense that the observations belonging to them are more homogeneous. Tree-growing is a recursive process. In the next step each descendant node is considered likewise, and new nodes are added by the same procedure. In principle this tree-growing process can lead to perfect purification (where each final node contains objects belonging to the same class), but, in practice, researchers apply some stopping criterion when, for instance, the number of objects in the final nodes should not be below a certain threshold.

For the classification problem other impurity measures can be used, such as the Gini-index. Trees can be grown to continuous response variables (the regression tree). In that case the most usual is to measure the goodness of fit with the mean squared error metric, but tree-growing can accommodate other measures as well.

It is clear that at the end we find a fully grown tree (if there is no stopping criterion) which gives a perfect fit, and therefore would not be very useful for prediction (an obvious case of overfitting). Still tree growing provides much information since the path to the full-grown tree is also important, it shows an optimal way to reach that. As usual overfitting leads to high variance, and it must be controlled. To make tree-growing a successful predictive device the bias-variance trade-off must be dealt with. Different approaches have been developed to use trees to get a prediction that is validated.

A mathematical description looks like this.

Every node A represents a subset of observations. The root node (R) contains all observations. Then

$$p_{iR} = \frac{n_i}{n}, i = 1, ..K$$

is the *a priori* probability (relative frequency) of class i in the sample.

The loss function in case of classification may be

$$\begin{aligned} L(i,j) &= 0, i = j, i, j = 1, ...K \\ L(i,j) &= 1, otherwise. \end{aligned}$$

Then the true class of observation x_h

$$\tau(x_h), \tau : X \to \{1, ...K\}$$

The relative frequency (probability) of A can be defined as

$$P(A) = \frac{n_A}{n}.$$

While the relative frequency of type i in A is

$$p_{iA} = \frac{n_{iA}}{n_A}.$$

We must give a classification rule at node A:

$$\tau(A) = \arg\max_i p_{iA}.$$

Let us define the entropy impurity function as:

$$I(A) = -\sum_{i=1}^{K} p_{iA} \log p_{iA}$$

We say that node A has left and right descendants (A_L, A_R). At each step we choose descendants so as to minimize average impurity:

$$P(A_L)I(A_L) + P(A_R)I(A_R).$$

We grow a tree until all nodes (i.e. end nodes) are pure. The resulting tree is called T, leading to perfect training sample classification. (The training error is 0.) We have every reason to think that it "overfits".

In case of a quantitative target variable the loss function is mean squared error, the impurity is average squared error at each node, and the estimate is the average of the target values belonging to the node in question.

Pruning the tree T The tree built by the above manner can be regarded as a non-parametric estimate of a two-valued function, where the procedure divides the input space into mutually exclusive regions, and assigns each observation to one of the classes depending on the region (final node) it belongs to. An alternative interpretation assigns a probability based on the relative frequencies of the corresponding region (final node), when the final nodes are not completely pure. There exists general theorems that assert that with a very large number of observations this estimate can be considered unbiased. However, it is also recognized that a very large (finely tuned) tree probably overfits (i.e. accommodates noise), resulting in reduced predictive abilities. Therefore, CART prunes the initially built tree using complexity cost pruning. In the first step of pruning one finds the best subtree, in the sense of least entropy or impurity, for a number of complexity classes, where a tree is more complex if it has more leaves. Then a validation procedure compares the best subtrees´ generalization capabilities, and the one with the best predictive score is chosen as the end product of the

procedure. (Concrete implementations may differ in the choice of complexity cost, and in the validation procedure.)

More formally, the perturbed loss function for a tree T_d is

$$\sum_{h=1}^{n} L(\tau(x_h), \tau(T_d(x_h)) + \omega \, |T_d|$$

where $T_d(x_h)$ is the end node in tree T_d where x_h belongs to, and $|T_d|$ is the cardinality of the end nodes in T_d.

For $\omega = 0$ T is the optimal (minimal loss) tree obviously. By increasing ω shorter and shorter trees become optimal. It can be proved that the process results in a series of sub-trees. If ω is infinite we get the root-tree as the optimal one.

To summarize: we get $0 = \omega_0 < \omega_1 < \omega_1 < ...\omega_M = \infty$, and for each interval $[\omega_i, \omega_{i+1}]$ there is an optimal tree of a certain size. How to choose the optimal sub-tree , i.e. the optimal ω (complexity cost)?

Determine βs as follows:

$$\begin{aligned} \beta_1 &= 0 \\ \beta_2 &= \sqrt{\omega_1 \omega_2} \\ &... \\ \beta_M &= \infty \end{aligned}$$

Do K-fold cross validation where for each n-$\frac{n}{K}$subsample estimate M models, one for each β. Compute the loss from the classifications and average it over the K subsample. You get a loss (sometimes called risk) for each β_j. Choose the β_j with the smallest error.

Interpretation of a tree The final tree can be interpreted as a decision tree where at each node some binary decision is made, leading to final decisions concerning where to classify a certain object. For any new observation one has to find its region in the input space, and make the corresponding classification as a prediction. (The alternative interpretation again is a probabilistic judgment, rather than a "yes-no" decision.) For regression trees the prediction equals the average at each node, thus it is basically a step function.

When interpreting the winning tree informally one can say that it suggests that important variables are those that have many and closer to the root splits in them, but researchers have also developed formal indicators to measure the relative importance of explanatory variables, based on the entropy reduction work they do.

Another possible use of CART models is by varying the input space: we can include (suspect) variables (either deemed as relevant or irrelevant), and see how they appear in the best decision tree. We can adapt the idea of Granger-causality as well: does the inclusion of a variable significantly improve the predictive performance of the model or not? As the CART algorithm does not lead

automatically to a better in-sample fit, after adding a new variable this question can (sometimes) be evaluated in a two-valued logic context, in contrast to Granger-causality where the measure of significance depends on the validity of maintained probabilistic assumptions.

Finally, CART algorithms can be used for "audience segmentation", as they are used in public health applications. One can identify non-trivial segments of society by their behaviour, enabling policy makers to adjust interventions targeted to these different groups.

5.3.3 Tree-based ensemble methods

Boosting Boosting is a tree-based algorithm where pruning is missing. We first grow a small tree, then model residuals with small trees consecutively, and finally add the models up.

Random forests CART is a greedy algorithm, as it strives at each step to achieve maximal purity increase. This results in higher variance, and instability (small changes in samples lead to large changes in the tree). Bagging is a version that addresses this problem by growing many trees, but on different bootstrap samples. Bootstrapping can be regarded as an alternative way of validation, and accordingly bagging does not use pruning, rather it averages over many large and unpruned trees.

A Random Forest is also constructed from a collection of trees, where the number of trees is a parameter set by the researcher. The prediction (estimate) a Random Forest regression gives is the average of the constituent trees' predictions. RF improves on bagging by randomizing variable choice at each cut-point, at each node only a random subset of explanatory variables are considered for a split. The cardinality of that subset is another parameter of the algorithm.

The main advantage of Random Forests is that the random and restricted manner of splitting achieves de-correlation among the many trees, while unbiasedness is not jeopardized. It has been argued that Random Forest regression is similar to other traditional non-parametric regression methods (e.g. k-nearest-neighbor algorithms), as it delivers some weighted average of "nearby" points as the prediction, when both the weights and the "nearby-ness" are determined in a data-driven way. All in all, with the presence of significant non-linearities, and with a relative abundance of explanatory variables Random Forest seems to be a successful and well-attested predictive methodology.

Though an outstanding method for prediction Random Forest regression has a problem: the results are not easily interpretable variable-wise. The demand for assessing the separate role of variables (their individual explanatory power) led to the proposal of several variable importance measures. As trees are grown from bootstrap samples a number of out-of-the-bag (OOB) observations belong to each tree, namely those data points that are not included in the sample for that particular tree. One can then calculate the prediction MSE for OOB data for each tree. Now the idea is that if a variable is unimportant it does not matter whether the predictions are generated with the help of their true

values, or are calculated from a random permutation of the true data. (The permutation shuffles only the values of the variable in question.) Then one can calculate the difference between the true and the permuted MSE, which must be small if the variable is unimportant. By averaging all such differences over all trees one obtains a measure of variable importance. This measure is obviously ad hoc.

5.3.4 Support vector machines (SVM)

SVM was originally developed for binary classification with 1 or -1 output.

The maximum margin classifier The maximum margin classifier is the "grandfather" of the SVM. It tries to separate two sets with hyperplanes, and tries to find the hyperplane that achieves maximal separation. It solves the following problem:

$$\begin{aligned} & \max_{\boldsymbol{\beta}} M \\ \sum \beta_i^2 &= 1 \\ y_i \boldsymbol{\beta}' \mathbf{x} &\geq M. \end{aligned}$$

This is a linear quadratic problem, for which there may not be any feasible solution. Obviously not any two sets are linearly separable.

The support vector classifier This is the immediate predecessor of the SVM. It relaxes the conditions of the maximum margin classifier, allowing that some observations be placed on the "wrong" side. The mathematical problem:

$$\begin{aligned} & \max_{\boldsymbol{\beta}} M \\ \sum \beta_i^2 &= 1 \\ y_i \boldsymbol{\beta}' \mathbf{x} &\geq M(1 - \epsilon_i) \\ \epsilon_i &\geq 0, \sum \epsilon_i \leq C. \end{aligned}$$

In fact an observation can be on the wrong side of the margin, or on the wrong side of the hyperplane. Here C is the tuning parameter ("budget"). It is possible that there is no feasible solution, but by increasing C we will get one sooner or later.

It can be proved that:

(1) the linear classifier can be written as

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle$$

where $\langle x, x_i \rangle$ is the inner product.

(2) To determine the unknown parameters we need only the $n(n-1)/2$ inner products of pairs of training observations.

(2) If S is the set of support points then

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i \langle x, x_i \rangle .$$

An observation that lies strictly inside the margin does not modify the classifier.

Support vector machines This is another way to solve the infeasibility problem. We enlarge the feature space with nonlinar functions of the original features. Then the classifier is linear in the modified feature space, though it corresponds to a nonlinear decision boundary in the original feature space. The basic idea is that we replace the inner product with its generalization, a positive definite kernel function (it is called the kernel trick):

$$K(x_i, x_j) \geq 0, symmetric.$$

Then the separating "plane" can be written as

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i K(x_i, x_j).$$

Several kernels have been proposed in the literature. A frequently used kernel is the following:

$$K(x_i, x_j) = \exp(-\gamma \sum_{k=1}^{p} (x_{ik} - x_{jk})^2).$$

The implicit enlarged feature space can be very high dimensional (even infinite), but we do not need to "enter" into it.

Here the tuning parameters are C and some parameter of the kernel (for instance γ above).

SVM and multiple classes There are several ways to extend SVM to multiple class classification.

(1) One versus one classification. In that case all possible K(K-1)/2 "matches" are played (binary classifications done), and the final classification is made by majority voting.

(2) One versus all classification. Here in each "match" one class plays against the rest, thus only K binary classification must be accomplished. In each of them belonging to the stand-alone group is coded as 1. The final classification of $\mathbf{x}$ occurs by argmaxing.

SVM can be extended to regression problems, too.

5.4 Literature

Hastie, T., & Tibshirani, R. & Friedman, J.(2008). The Elements of Statistical Learning; Data Mining, Inference and Prediction.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: Springer.

6 Time series analysis

6.1 The stochastic theory of time series

A discrete time stochastic process is the ensemble of (countably) infinitely many ordered random (vector) variables. The elementary event is an infinite trajectory. If all the finite dimensional marginal distributions exist then the whole process constitutes a regular probability space.

The basic ingredient of time series is the white noise (a sort of "unit vector"):

$$\begin{aligned} E(\epsilon_t) &= 0 \\ E(\epsilon_t \epsilon_{t'}) &= 0, t \neq t' \\ var(\epsilon_t) &< \infty. \end{aligned}$$

The Gaussian white noise is a normally distributed white noise.

Second order moments for stochastic processes are autocovariance and autocorrelations functions:

$$cov(x_t, x_{t-k}) = E(x_t x_{t-k}) - E(x_t) E\left(x_{t-k}\right)$$

$k = ..., -1, 0, 1, ...$

$$cor(x_t, x_{t-k}) = \frac{cov(x_t, x_{t-k})}{\sqrt{var(x_t) var(x_{t-k})}}.$$

If we have a vector valued stochastic process than cross-autocovariances can be defined as:

$$cov(x_t, y_{t-k}) = E(x_t y_{t-k}) - E(x_t) E\left(y_{t-k}\right), k = ..., -1, 0, 1, ...$$

for each x, y pair.

6.1.1 An important subclass: stationary stochastic processes

From now on we will focus on single variable (one-dimensional) series. Strong stationarity of time series means that distribution functions depend only on distances.

For all k, τ and $t_1, ...t_k$:

$$F_X(x_{t_1+\tau}, ...x_{t_k+\tau}) = F_X(x_{t_1}, ...x_{t_k}),$$

thus the joint distributions are not functions of time.

This obviously imples that

$$F_X(x_{t_1+\tau}) = F_X(x_{t_1}),$$

i.e. the same unconditional distribution describes x at each time.

Weak stationarity means that the first and second moments depend only on distance:

$$E(x_{t+\tau}) = E(x_t),$$

$$cov(x_{t+\tau}, x_{t-k+\tau}) = cov(x_t, x_{t-k}).$$

In general strong stationarity implies weak stationarity, and for a Gaussian process, weak stationarity implies strong stationarity, too.

It is an important observation that linear combination of stationary series are stationary. With stationary series we can talk of kth order autocovariance matrices unequivocally, which are symmetric and positive definite. For instance

$$\begin{bmatrix} var(x_t) & cov(x_t, x_{t+1}) \\ cov(x_{t+1}, x_{t+2}) & var(x_{t+1}) \end{bmatrix}.$$

is symmetric, $cov(x_{t-1}, x_t) = cov(x_t, x_{t+1})$, moreover $var(x_t) = var(x_{t+1})$.

The traditional notation is γ_0 for the variance, and $\gamma_{-k} = \gamma_k$ for the kth autocovariance.

An important subclass of stationary series are the mean ergodic processes. A mean ergodic process is such that $\frac{\sum_{i=0}^{T} X_{t+i}}{T}$

converges as $T \to \infty$ in squared mean to μ, the unconditional mean of the process. In other words, averaging over any trajectory we obtain a consistent estimate of the mean. It can be proved that

$$\sum_{i=0}^{\infty} abs(\gamma_i) < \infty$$

is a sufficient condition for ergodicity.

In the following we will usually deal with ergodic processes,

6.1.2 Representation in the time domain of covariance-stationary processes

A fundamental theorem (The Wold Representation Theorem) asserts that for covariance-stationary processes the mean can be regarded as a deterministic function of time (sometimes called a signal) plus a purely indeterministic process, i.e. an infinite linear combination of white noise. In formulas:

$$x_t = f(t) + \sum_{i=0}^{\infty} a_i \varepsilon_{t-i}, a_0 = 1,$$

where $f(t)$ deterministic, ε_t white noise, moreover $\sum_{i=0}^{\infty} a_i^2 < \infty$.

The white noise process is not "noise" in the sense this word is used in the literature. ε_t is called innovation (or shock). It is also the CEF residual when the condition is the whole history of the process up to $t-1$:

$$x_t = E_{t-1}(x_t) + \epsilon_t.$$

(The meaning of E_t: expected value conditioned on everything occurring before and including time t.) Shocks are "unexpected", but have an influence beyond the time they occur, while "noise" is usually taken to be something that has no influence on the future course of the process, it is merely a matter of imperfect observation.

6.2 Mathematical detour

6.2.1 Stability of linear difference equations

Consider the following equation:

$$\pi_n = \alpha_1 \pi_{n-1} + - - + \alpha_p \pi_{n-p}.$$

Let λ_i be a root of the following polinomial:

$$\lambda^p - \alpha_1 \lambda^{p-1} - .. - \alpha_p = 0.$$

Then for any A_i

$$\pi_n = A_i \lambda_i^n$$

is a solution of the difference equation.
Every solution can be written as

$$\pi_n = \sum_i^p A_i \lambda_i^n.$$

The p initial conditions determine the A_i constants.
Obviously π_n converges if and only if all λs are less than 1 in absolute value.

6.2.2 A useful tool: lag polinomials

A useful mathematical tool in times series analysis is the lag operator:

$$Lx_t = x_{t-1},$$

that translates any series one step back in time.
The powers of L can be defined naturally as

$$L^n x_t = L(L^{n-1})x_t = x_{t-n},$$

where

$$L^0 x_t = x_t.$$

Obviously there exists L^{-1} for which

$$\begin{aligned} L^{-1}(Lx_t) &= x_t \\ L^{-1}x_t &= x_{t+1.} \end{aligned}$$

L^{-1} is called the forward operator, and sometimes is denoted by F.
Finite order polinomials of L can be defined accordingly as

$$A(L)x_t = \sum_{i=-k}^{k} \alpha_i(L^i x_t).$$

With traditional notation a pth order linear difference equation can be written as:

$$x_t = a_1 x_{t-1} + ... + a_p x_{t-p} + f_t,$$

where f_t is called an exogenous forcing process. Then in terms of lag operators the equation can be rewritten as

$$(1 - \sum_{i=1}^{p} \alpha_i L^i)x_t = A(L)x_t = f_t, \alpha_p \neq 0.$$

One can guess formally that the solution can be written a:

$$x_t = A(L)^{-1}\epsilon_t.$$

The question is whether we can assign a precise menaing to the inverse $A(L)^{-1}$ as a lag operator. If we want to have a causal solution (where the present does not depend on the future) $A(L)^{-1}$ should contain only positive powers of L. The question boiles down to the problem whether we can find parameters $\pi_0, \pi_1, ...\pi_n, ...$ that satisfy:

$$(L^0 - \alpha_1 L - ...\alpha_p L^p)(\pi_0 L^0 + \pi_1 L + ...\pi_n L^n + ...) = (L^0 + 0L + ... + 0L^n + ...).$$

In fact we also require that the $\sum \pi_i^2$ series converge to 0. It can be proved that

$$\pi_n = -(a_1 \pi_{n-1} + - - + a_m \pi_{n-m}),$$

in other words the parameters of the inverse operator satisfy this linear difference equation.

We know that for π_n to converge it is necessary and sufficient that the roots of

$$\lambda^m + a_1 \lambda^{m-1} + .. + a_m$$

be within the unit circle, therefore the

$$1 + b_1 L^1 + .. + b_m L^m$$

polinomial must have all roots outside the unique circle.
Let us illustrate this important point with the simplest case. Suppose

$$x_t = \alpha x_{t-1}.$$

Then convergence requires that $abs(\alpha) < 1$. The lag polinomial form of the equation is

$$(1 - \alpha L)x_t = 0.$$

Clearly the root of $1 - \alpha L$ is $\frac{1}{\alpha}$.

6.3 ARMA processes: making the Wold Theorem practical

The Wold Theorem states that stationary series can be characterized by an infinite series of parameters having a time distance effect interpretation. Finite data sets demand that we have models with a finite number of parameters. ARMA processes are an "empirically realizable" subset of stationary processes. First we introduce the pure AR, and then the pure MA processes, before defining the general ARMA process.

6.3.1 AR (p) processes

The general $AR(p)$ process can be defined as

$$x_t = C + \alpha_1 x_{t-1} + ... + \alpha_p x_{t-p} + \epsilon_t.$$

Let $\mu = E(x_t)$.Then

$$\begin{aligned} \mu &= C + (\alpha_1 + ... + \alpha_p)\mu. \\ \mu &= \frac{C}{1 - \alpha_1 ... - \alpha_p}. \end{aligned}$$

If we change to the variable $y_t = x_t - \mu$, with mean 0, then

$$y_t = \alpha_1 y_{t-1} + ... + \alpha_p y_{t-p} + \epsilon_t.$$

For simplicity we will deal with zero mean series in the following.
There are several identical formulations:

$$\begin{aligned} x_t &= \sum_{i=1}^{p} \alpha_i x_{t-i} + \epsilon_t, \\ x_t &= (\sum_{i=1}^{p} \alpha_i L^P) x_{t-i} + \epsilon_t, \\ (1 - \alpha_1 L - \alpha_2 L^2 - ...\alpha_p L^P) x_t &= A(L) x_t = \epsilon_t, \end{aligned}$$

where ϵ_t white noise with variance σ^2.

We can determine x_t in the Wold-framework as

$$x_t = (1 - \alpha_1 L - \alpha_2 L^2 - ...\alpha_p L^P)^{-1} \epsilon_t = A(L)^{-1} \epsilon_t,$$

where $A(L)^{-1}$ is an infinite lag polinom, provided that the roots of the polinomial

$$1 - \alpha_1 L - \alpha_2 L^2 - ...\alpha_p L^P$$

exceed 1 in absolute value. It is called the impulse response function.

The autocovariance function can be determined from the Yule-Walker equations, to be derived now.

Let us start from the

$$x_t = \sum_{i=1}^{p} \alpha_i x_{t-i} + \epsilon_t$$

expression. Multiplying both sides with ϵ_t, and taking expectations we get

$$E(x_t \varepsilon_t) = \sigma^2.$$

Then multiplying with x_t and taking expectation we obtain

$$\gamma_0 = \alpha_1 \gamma_1 + ...\alpha_p \gamma_p + \sigma^2.$$

Multiplying both sides with x_{t-k} $(k = 1, ...p)$, and again taking expectations the new result is

$$\gamma_k = \alpha_1 \gamma_{k-1} + ...\alpha_p \gamma_{k-p},$$

where, because of stationarity,

$$\gamma_{k-p} = \gamma_{p-k}.$$

This is a linear system of $p + 1$ equations in as many variables, from which one can determine the unknown $\gamma_0, \gamma_1, ...\gamma_p$. Then autocorrelations (ρ_k) are obtained by dividing by γ_0.

For $k > p$ autocovariances satisfy the following difference equation:

$$\gamma_k = \sum_{i=1}^{k-1} \alpha_i \gamma_{k-i}.$$

Equivalently for autocorrelations

$$\rho_k = \sum_{i=1}^{k-1} \alpha_i \rho_{k-i}.$$

These difference equations can be uniquely solved taking into account the $p+1$ initial values, computed in the Yule-Walker equations. A necessary condition for (ergodic) stationarity is that this equation be asymptotically stable, in other words autocovariances converge to 0.

6.3.2 MA (q) processes

The general $MA(q)$ process can be defined as

$$x_t = C + u_t + \beta_1 u_{t-1} + ... + \beta_q u_{t-q}.$$

Then

$$E(x_t) = C,$$

and if $y_t = x_t - C$, then

$$y_t = u_t + \beta_1 u_{t-1} + ... + \beta_q u_{t-q}$$

is a 0 mean $MA(q)$ process. Again we restrict our attention to zero mean processes.

A zero-mean $MA\ (q)$ process can be defined as

$$\begin{aligned} x_t &= \varepsilon_t + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i}, \\ x_t &= (1 - \sum_{i=1}^{q} \beta_i L^i)\varepsilon_t. = B(L)\varepsilon_t. \end{aligned}$$

Autocovariances vanish after q periods:

$$\begin{aligned} \gamma_0 &= \sigma^2(1 + \sum_{i=1}^{q} \beta_i^2), \\ \gamma_k &= \sigma^2(\beta_k + \sum_{i=1}^{q-k} \beta_i \beta_{i+k}), k = 1, ...q, \\ \gamma_k &= 0, k > q. \end{aligned}$$

The relationship between parameters and autocovariances is non-linear (quadratic), and therefore there are multiple solutions. Whenever $B(L)^{-1}$ exists (invertibility) :

$$\varepsilon_t = B(L)^{-1} x_t.$$

Because autocovariances vanish there are always invertible representations, but also non-invertible ones, because of multiplicity.

6.3.3 Generalization: ARMA (p,q) with non-zero mean

$$\begin{aligned} A(L)x_t &= C + B(L)\epsilon_t, \\ \mu &= (1 - \alpha_1 - ... - \alpha_p)^{-1} C \\ y_t &= x_t - \mu \\ A(L)y_t &= B(L)\epsilon_t. \end{aligned}$$

where $A(L)$ and $B(L)$ are finite lag polinomials, and ϵ_t white noise. Then in the case of stationarity:

$$x_t = A(L)^{-1} B(L)\epsilon_t,$$

which is a Wold-representation (an infinite MA representation). This is also called the impulse response.

If $B(L)^{-1}$exists

$$B(L)^{-1} A(L) x_t = \epsilon_t,$$

the process has an infinite AR representation, and is called invertible.

6.3.4 Partial autocorrelation in the stationary case

We defined the linear projection of y on $(x_1, x_2 ... x_n)$ with the following expressions:

$$\begin{aligned} \overline{y} &= \boldsymbol{\beta}\mathbf{x}, \\ cov(\overline{y}, \mathbf{x}) &= \mathbf{0}. \end{aligned}$$

Let $\widehat{x_{-i}}$ be the projection of x_i on x_{-i} (e obtain x_{-i} from x by skipping x_i). Let

$$\widetilde{x_{-i}} = \widehat{x_{-i}} - x_i,$$

be the projection error. Then partial covariances (correlations) are defined as:

$$\begin{aligned} pcov_{x_{-i}}(x_i, x_j) &= cov(\widetilde{x_{-i}}, \widetilde{x_{-j}}) \\ pcor_{x_{-i}}(x_i, x_j) &= cor(\widetilde{x_{-i}}, \widetilde{x_{-j}}). \end{aligned}$$

Partial autocovariances (autocorrelations) are defined as:

$$
\begin{aligned}
pacov_k(x_t, x_{t-k}) &= pcov_{x_{t-1}, x_{t-2}, \ldots x_{t-k+1}}(x_t, x_{t-k}) \\
pacor_k(x_t, x_{t-k}) &= pcor_{x_{t-1}, x_{t-2}, \ldots x_{t-k+1}}(x_t, x_{t-k})
\end{aligned}
$$

Thus all the observations between t and $t-k$ are partialled out. Notice that the partial correlation between two variables depends on the conditioning variables, thus it is not a unique number. However, the definition of partial autocorrelation assigns a unique number as the conditioning is determined unequivocally.

6.3.5 The statistical approach: Box-Jenkins analysis

Identification ARMA models have particular shapes for the auto and partial autocorrelation functions, depending on p and q. The identification phase consists in estimating these functions, and guessing at p and q from the estimates.

The sample mean, the sample autocovariance and autocorrelation are consistent estimators in the ergodic case:

$$\overline{x} = \frac{1}{T} \sum x_t,$$

$$acov_0^s = \frac{1}{T} \sum (x_t - \overline{x})^2,$$

$$acov_k^s = \frac{1}{T} \sum_{t=1}^{T-k} (x_t - \overline{x})(x_{t+k} - \overline{x}),$$

$$acor_k^s = \frac{acov_k^s}{acov_0^s}.$$

If the process is white noise the asymptotic distribution of the sample autocorrelations is normal with $1/T$ variance. From this one can calculate confidence intervals.

The Box-Pierce statistics is used to test the null-hypothesis that m autocorrelations are 0 :

$$Q_{BP} = T \sum_{k=1}^{m} r_k^2.$$

This statistics is asymptotically χ^2.

Partial autocovariances are estimated from the autoregression coefficients. Let b_k be the

$$\widehat{x_t} = \sum_{i=1}^{k} b_i x_{t-it}$$

last coefficient in this empirical projection. As

$$var(\widetilde{x_{t-k}}) = var(\widetilde{x_t})$$

therefore

$$b_k = \frac{cov(\widetilde{x_t}, \widetilde{x_{t-k}})}{var(\widetilde{x_{t-k}})} \frac{\sqrt{var(\widetilde{x_{t-k}})}}{\sqrt{var(\widetilde{x_t})}} = \rho_k.$$

Estimation of ARMA processes Pure AR processes can be consitently estimated by OLS. Otherwise the two most frequent methods are conditional least squares and maximum likelihood. The former's advantage is that it can dispose of a specific distributional assumption.

Conditional least squares We write down recursively the residuals as functions of parameters and observables, and then minimize the squared residuals. This method can be illustrated by a simple example.

An example: conditional least squares estimation of ARMA (1,1) Here the residual is:

$$u_t = x_t - \alpha x_{t-1} - \beta u_{t-1}.$$

The least squares problem:

$$\min_{\alpha,\beta} \sum_{t=2}^{T} u_t^2.$$

We can start from $t = 2$, thus x_1 must be a condition. However we still need u_1. The simplest assumption is that $u_1 = 0$ (equals its expected value). Except for $\beta = 0$ this is a nonlinear optimization problem.

Maximum Likelihood estimation The novelty of time series models, with respect to i.i.d. samples, is that the observations are mutually dependent. If we assume that the sample has a multidimensional (centralized) normal distribution then the density is

$$F(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp(-\frac{\mathbf{y}'\Sigma^{-1}\mathbf{y}}{2}).$$

The rules of conditional probability tell us that

$$f(x, y) = f(x \mid y) f(y),$$

and this property can be explored to write down the likelihood function in specific cases. For simplicity consider the AR(1) model.

$$
\begin{aligned}
f(y_1, ..y_T) &= f_{y_T|y_{T-1},...y_1} * f(y_1, ..y_{T-1}) \\
f(y_1, ..y_{T-1}) &= f_{y_{T-1}|y_{T-2,...}y_1} * f(y_1, ..y_{T-2}) \\
&... \\
f(y_1, ..y_T) &= f_{y_1} * f_{y_2|y_1} * .. * .f_{y_{T-1}|y_{T-2,...}y_1} * f_{y_T|y_{T-1},...y_1}.
\end{aligned}
$$

The conditional distrubutions are known as

$$y_t \mid y_{t-1}, .. \sim N(\alpha y_{t-1}, \sigma^2).$$

We have no information on the conditional distribution of y_1. However, we "know" the unconditional distrubution of y_1 :

$$y_1 \sim N(0, \frac{1}{1-\alpha^2}\sigma^2).$$

Plugging this into the formula above the product of the distributions provides the likelihood to be maximized as a function of α and σ. The formula can be generalized to AR(p). If we have MA terms the expression is much more complicated.

Selecting the best model Normally, after the identification phase we have several candidate models. Each of them is estimated, then diagnostic testing is applied, and those models are preferred that show favourably diagnostic properties. (For instance residuals appear to be normal, and they do not appear to be autocorrelated etc..) Also, it is customary to compute information criteria for selecting the best model. Out of sample forecasting exercises and tests are not frequent in econometrics, but potentially these would provide the best solution.

6.4 Some generalizations of ARMA in the time domain

6.4.1 A non-stationary generalization: ARIMA (p,d,q)

In that case the autocorrelation function suggests that the series is non-stationary. Frequently, after one or two-differencing ($\nabla x_t = x_t - x_{t-1}, \nabla^2 x_t = \nabla x_t - \nabla x_{t-1}$) the resulting series can be taken as stationary. With the lag operator differencing can be written as: $(1-L)^d x_t = \Delta^d x_t$.

Then

$$A(L)(1-L)^d x_t = B(L)u_t.$$

is called an $ARIMA(p,d,q)$. Its speciality is that the AR polynomial contains unit roots. The usual treatment of ARIMA models is that after the necessary differencing the model is treated as an $ARMA(p,q)$. Naturally forecasting of the undifferenced series must take this into account.

There exist tests for deciding whether differencing results in a stationary process. In these tests usually the null hípothesis is unit root in the process,

in other words that that the autoregressive polinomial has a unit root. For example in the case of the simple Dickey-Fuller test one estimates the following equation by OLS,

$$x_t = \alpha + \beta x_{t-1} + At + u_t.$$

and then tests whether $\beta = 1$. However, the distribution of the "t statistic" calculated from the OLS estimate is different from the t distribution, therefore using the test requires specific tables for the corresponding true distribution.

Trend-stationary and difference-stationary processes Suppose that

$$x_t = At + u_t.$$

where u_t is a stationary ARMA. Then

$$x_t - x_{t-1} = A + u_t - u_{t-1} = A + (1 - L)u_t.$$

Therefore x_t is $I(1)$, but it is not invertible, since there is a unit root in the MA polinomial. These processes are called trend-stationary, since we would get a stationary process after subtracting the trend. The correct treatment of the process would involve the simultaneous estimation of the trend and the residual process.

It has become almost an article of faith among macroeconomists that macroeconomic time series can be stationarized by differencing. In the next section we explore models that rely on this assumption, but in a multiple time series context.

6.4.2 Seasonally integrated series

Another generalization makes the assumption that

$$(1 - L)^d (1 - L^S)^D x_t$$

is stationary. In this case there are seasonal unit roots in the lag polinomial. Econometricians usually prefer to work with seasonally adjusted data, and the treatment of seasonality belongs to data pre-processing.

6.4.3 Fractionally integrated series

We have a formal generalization of differencing:

$$(1 - L)^d x_t = \epsilon_t.$$

What is the meaning of d, when it is any real number? Take the power series expansion of

$$(1 - L)^d$$

around $L = 0$. This gives

$$(1-L)^d = 1 - dL - \frac{d(d-1)}{2!}L^2 - \frac{d(d-1)(d-2)}{3!}L^3 - ...$$

This is an infinite lag polinomial. The coefficients satisfy:

$$\phi_j = \frac{j-1-d}{j}\phi_{j-1}.$$

If $d < 0.5$ then x_t is stationary, but not absolute summable. Therefore this is called a long-memory process. Certain economic time series are supposed to be well described by long-memory processes, but their identification needs long data series.

6.4.4 ARCH and its generalizations

Financial time series often exhibit heteroskedasticity in time. The theoretical counterparts are autoregressive conditional heteroskedasticity (ARCH) models. The simplest one perhaps:

$$z_t = \rho z_{t-1} + u_t,$$

$$E(u_t \mid u_{t-1,...})) = 0,$$

$$h_t = var(u_t \mid u_{t-1,...})) = \omega_0 + \sum \omega_i u_{t-i}^2.$$

A generalization of ARCH is GARCH (generalized ARCH), where

$$h_t = var(u_t \mid u_{t-1,...})) = \omega_0 + \sum \omega_i u_{t-i}^2 + \sum \psi_i h_{t-i}^2.$$

This model produces distributions with fatter tails than the normal's, another feature of many financial time series. A number of variations and further generalizations have been developed in the financial econometrics literature. The identification and estimation of these models is usually based on high frequency data that are not available for macroeconomists.

6.5 Multiple time series analysis in the time domain

Economists consider usually several time series simultaneously. The one-dimensional time-domain theory can be extended to the multiple dimension case. The simplest is the extension of the $AR(p)$ model to multiple time series.

6.5.1 VAR representation

Analogously to the one-dimensional case we say that the following is a $VAR(p)$ model, where VAR stands for vector autoregression.

$$\mathbf{x}_t = \mathbf{A}_1\mathbf{x}_{t-1} + ... + \mathbf{A}_p\mathbf{x}_{t-p} + \boldsymbol{\epsilon}_t,$$

where $\mathbf{x}_t$ has n components, and $\boldsymbol{\epsilon}_t$ is vector white noise with Ω positive definite instantaneous covariance matrix. In general we assume that Ω is not diagonal, thus there is contemporaneous connection between the different elements of the $\mathbf{x}_t$ vector.

The lag polinomial form analogously is:

$$(\mathbf{I} - \mathbf{A}_1 L - ... - \mathbf{A}_p L^p)\mathbf{x}_t = \mathbf{A}(L)\mathbf{x}_t = \boldsymbol{\epsilon}_t,$$

where, for example, $\mathbf{A}_1 L$ is a nxn matrix whose each element is multiplied by L (symbolically).

It turns out that through redefining variables an nxn $VAR(p)$ is mathematically equivalent with a one-variable $AR(pxn)$ model. Therefore mathematical results can be transported. These yield a condition for stationarity: the determinant

$$\det(\mathbf{I} - \mathbf{A}_1 L - ... - \mathbf{A}_p L^p),$$

which is a pxn polinomial in L, must have roots exceeding 1 in absolute value.Though VMA (vector moving average) and $VARMA$ (vector autoregression and moving average) models can also be defined, they are not used frequently.

Impulse response function The $MA(\infty)$ form gave the impact of innovations (shocks) on different horizons in the one-variable case. Here there is an analogous definition. The $VMA(\infty)$ form:

$$\mathbf{x}_t = (\mathbf{I} - \mathbf{A}_1 L - ... \mathbf{A}_p L^p)^{-1}\boldsymbol{\epsilon}_t$$

$$(\mathbf{I} - \mathbf{A}_1 L - ... \mathbf{A}_p L^p)^{-1} = I + \Pi_1 L + \Pi_2 L^2 + ... + \Pi_k L^k + ...$$

$$\mathbf{x}_t = \epsilon_t + \Pi_1\boldsymbol{\epsilon}_{t-1} + \Pi_2\boldsymbol{\epsilon}_{t-2} + ... + \Pi_k\boldsymbol{\epsilon}_{t-k} + ...$$

is called the impulse response function. The interpretation is that the $\Pi_k^{(ij)}$($\Pi_k^{(ij)} = \frac{\partial x_t^{(i)}}{\partial \epsilon_{t-k}^{(j)}}$) element is the marginal effect of shock j on variable x_i after k periods.

The matrix of long-run coefficients is an interesting analytic tool as it gives the cumulative (on an infinite horizon) effect of shocks:

$$(\mathbf{I} - \mathbf{A}_1 - ... \mathbf{A}_p)^{-1} = \mathbf{\Pi},$$

where $\Pi^{(ij)} = \lim \sum_{k=0}^{\infty} \frac{\partial x_t^{(i)}}{\partial \epsilon_{t-k}^{(j)}}$.

However, there is a problem with the interpretation: as $\mathbf{\Omega}$ is non-diagonal the different components of ϵ do not vary independently, therefore the partial derivatives cannot be interpreted unequivocally. Econometricians found a simple "solution" for this problem: let us suppose that there exist "fundamental" shocks ($\mathbf{u}$) with diagonal covariance matrix, and the VAR shocks ($\boldsymbol{\epsilon}$), are linear combinations of them:

$$\boldsymbol{\epsilon} = \mathbf{Q}\mathbf{u}.$$

It follows that

$$cov(\mathbf{u}) = \mathbf{\Omega}_u = \mathbf{Q}^{-1}\mathbf{\Omega}(\mathbf{Q}^{-1})'.$$

There are infinitely many $\mathbf{Q}$ with the property that $\mathbf{\Omega}_u$ diagonal. In that case $\mathbf{Q}$ can be written as

$$\mathbf{Q} = \mathbf{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sum_{i=1}^{n} \mathbf{q}_i \mathbf{q}_i'.$$

The modified impulse response function in terms of $\mathbf{u}$ is

$$\mathbf{x}_t = (\mathbf{I} - \mathbf{A}_1\mathbf{L} - ...\mathbf{A}_p\mathbf{L}^p)^{-1}\mathbf{Q}\mathbf{u}_t.$$

If we make enough assumptions to achieve uniqueness we obtain what is called the SVAR (structural VAR) analysis.

The easiest choice is if we assume that $\mathbf{Q}$ is lower triangular (Cholesky-decomposition) which can be interpreted as the existence of a causal chain within a period. For instance it is frequently assumed that prices do not react quickly to changes in supply, in this sense within a quarter prices affect demand or supply, but not vice versa.

Another popular approach is making long-run restrictions on the $(\mathbf{I} - \mathbf{A}_1\mathbf{L} - ...\mathbf{A}_p\mathbf{L}^p)^{-1}\mathbf{Q}$, which is the matrix of the long-run effects in the transformed model.

Variance-decomposition Variance decomposition stands for decomposing the mean (squared) prediction error due to different shocks at different horizons. It is meaningful if we have structural (orthogonal) shocks, only. The mathematical derivation is the following:

$$\begin{aligned}
x_{t+s} &= \boldsymbol{\epsilon}_{t+s} + \mathbf{\Pi}_1\boldsymbol{\epsilon}_{t+s-1} + ... + \mathbf{\Pi}_{s-1}\boldsymbol{\epsilon}_{t+1} + \mathbf{\Pi}_s\boldsymbol{\epsilon}_t + \mathbf{\Pi}_{s+1}\boldsymbol{\epsilon}_{t-1}... \\
\widetilde{x_{t+s,t}} &= \mathbf{\Pi}_s\boldsymbol{\epsilon}_t + \mathbf{\Pi}_{s+1}\boldsymbol{\epsilon}_{t-1} + ... \\
x_{t+s} - \widetilde{x_{t+s,t}} &= \boldsymbol{\epsilon}_{t+s} + \mathbf{\Pi}_1\boldsymbol{\epsilon}_{t+s-1} + ... + \mathbf{\Pi}_{s-1}\boldsymbol{\epsilon}_{t+1}.
\end{aligned}$$

where $\widetilde{x_{t+s,t}}$ is prediction made at t for s period ahead. Then:

$$
\begin{aligned}
E((\mathbf{x}_{t+s}-\widetilde{x_{t+s,t}})(\mathbf{x}_{t+s}-\widetilde{x_{t+s,t}})' &= \mathbf{\Omega}+\mathbf{\Pi}_1\mathbf{\Omega}\mathbf{\Pi}_1'+..+\mathbf{\Pi}_{s-1}\mathbf{\Omega}\mathbf{\Pi}_{s-1}' \\
&= \sum_{i=1}^{n}\left[\mathbf{I}\mathbf{q_i}\mathbf{q_i'}+\mathbf{\Pi}_1\mathbf{q}_i\mathbf{q}_i'\mathbf{\Pi}_1'+..+\mathbf{\Pi}_{s-1}\mathbf{q}_i\mathbf{q}_i'\mathbf{\Pi}_{s-1}'\right].
\end{aligned}
$$

The vector of mean squared prediction errors is the diagonal of this matrix. (The off-diagonal elements show covariances between predictions.) The ith innovation's share in the prediction error of the jth variable is:

$$
\frac{diag(\left[\mathbf{I}\mathbf{q_i}\mathbf{q_i'}+\mathbf{\Pi}_1\mathbf{q}_i\mathbf{q}_i'\mathbf{\Pi}_1'+..+\mathbf{\Pi}_{s-1}\mathbf{q}_i\mathbf{q}_i'\mathbf{\Pi}_{s-1}'\right])}{MSE_{j,s}}.
$$

6.5.2 Cointegration

Suppose in the one-variable case that a variable is either stationary or $I(1)$, difference stationary. (Of course there exist other possibilities, but we ignore them now.) In the multiple variable case it may happen that though all variables are $I(1)$, still there exist some linear combination of them which is stationary. Many macroeconomic time series look like $I(1)$ variables, but simple functions of them look rather stationary (for instance the share of consumption in GDP). Certain economic theories can be formulated as stationarity of functions of variables. It turns out that considering the possibilty of this feature of time series, called cointegration, results in differences for the properties of estimators, as well.

The case where $\mathbf{x}_t$ is I(1) Let

$$
\mathbf{x}_t=\mathbf{A}_1\mathbf{x}_{t-1}+\mathbf{A}_2\mathbf{x}_{t-2}+...\mathbf{A}_p\mathbf{x}_{t-p}+\boldsymbol{\epsilon}_t,
$$

be a $VAR(p)$, with all elements of $\mathbf{x}_t$ being $I(1)$ variables.
The equation can be equivalently rewritten as:

$$
\begin{aligned}
\boldsymbol{\nabla}\mathbf{x}_t &= (\mathbf{A}_1+\mathbf{A}_2+..+\mathbf{A}_p-\mathbf{I})\mathbf{x}_{t-1}-(\mathbf{A}_2+..\mathbf{A}_p)\boldsymbol{\nabla}\mathbf{x}_{t-1}+ \\
&\quad -(\mathbf{A}_3+...\mathbf{A}_p)\boldsymbol{\nabla}\mathbf{x}_{t-2}-...-\mathbf{A}_p\boldsymbol{\nabla}\mathbf{x}_{t-p+1}+\boldsymbol{\epsilon}_t.
\end{aligned}
$$

Granger's Representation Theorem Case 1: $\Pi=\mathbf{A}_1+\mathbf{A}_2+...\mathbf{A}_p-\mathbf{I}=\mathbf{0}$. It means that Π has n 0 eigenvalues. Then the VAR is stationary in differences.

Case 2: $dim(\Pi)=r, 0<r<n$. Π has $n-r$ 0 eigenvalues. Then there exist $\Gamma(nxr)$ and $\Psi(rxn)$ matrices for which

$$
\Gamma\Psi=\Pi
$$

and

$$
\Psi x
$$

is stationary, where the rows of Ψ are eigenvectors of Π belonging to the r non-zero eigenvalues.

(The possibility $dim(\Pi) = n$ is excluded by the I(1) assumption.)

There exist several methodologies to establish and estimate cointegration relationships. Johansen's method is relatively easily algorithmized.

Johansen's method 1. Estimate the regression

$$\nabla \mathbf{x}_t = \mathbf{\Pi} \mathbf{x}_{t-1} + \mathbf{\Phi}_1 \nabla \mathbf{x}_{t-1} + . - + \mathbf{\Phi}_{p-1} \nabla \mathbf{x}_{t-p+1} + \boldsymbol{\epsilon}_t .$$

2. Test the number of non-zero eigenvalues (r) in the estimated $\mathbf{\Pi}$ matrix.

3. If you accept the hypothesis that $r = 0$ then re-estimate the equation by setting $\Pi = 0$.

4. If you accept the hypothesis that $0 < r < n$, then calculate r eigenvectors of the estimated Π, corresponding to the largest r eigenvalues. Form the "cointegrating" relationships as the variables $\mathbf{z}_{t-1} = \widehat{\Psi} x_{t-1}$.

Then estimate the regression:

$$\nabla \mathbf{x}_t = \mathbf{\Gamma} \mathbf{z}_{t-1} + \mathbf{\Phi}_1 \nabla \mathbf{x}_{t-1} + ... + \mathbf{\Phi}_{p-1} \nabla \mathbf{x}_{t-p+1} + \boldsymbol{\epsilon}_t .$$

From the coefficients of this regression one can determine the coefficient estimates of the original model. (For example $\mathbf{\Phi}_1 = -(\mathbf{A}_2 + .. \mathbf{A}_p), \mathbf{\Phi}_2 = -(\mathbf{A}_3 + .. \mathbf{A}_p), ..., -\mathbf{A}_p = \mathbf{\Phi}_{p-1}$ etc.)

The important point is that if cointegration exists the estimation in differenced form is inconsistent, while the estimation in levels is not efficient.

Exogeneity and Granger-causality Let the target variable be y, and the explanatory variable(s) x. Multiple time series analysis traditionally is concerned with concepts of exogeneity and causality. Weak exogeneity means that the parameters of interest belonging to the conditional expectation function can be estimated by ML without knowing the parameters of the marginal processes of the corresponding variables. It results in efficient estimation, and it is usually easily assumed without much thinking.

Granger-causality is defined without proper respect to the traditional intuitive concept of causality. According to it x is not a Granger-cause of y if x does not improve the forecast error of y. This is frequently tested. If Granger-causality is not rejected in either direction then it gives an indication that a VAR must include both variable.

Strong exogeneity means that x is weakly exogenous plus y is not a Granger-cause of x. If we want to forecast y conditioned on x, then the fulfllment of this condition gives a sort of green light.

6.6 Signal processing and time series analysis

Signal processing is a problem when given finite observations $(g(t_i), i = 0, ...T)$ made on a function $f(t)$ we want to recover $f(t)$, where

$$g(t) = f(t) + \epsilon_t.$$

There is obvious similarity with the Wold Representation Theorem, the important difference is that ϵ_t is usually thought of as noise, rather than as a shock (or innovation) that drives the process.

6.6.1 General mathematical background

A usual assumption is that $f(t)$ belongs to a function space, normally the space of square integrable functions.

This space is linear and

$$\int f(t)g(t)dt$$

is a natural inner product. In inner product spaces angles, and orthogonality can be defined. Orthogonal functions are such that:

$$\int f(t)g(t)dt = 0.$$

6.6.2 Some general properties of inner product spaces

Orthogonal elements are linearly independent: $\langle u_i, u_j \rangle = 0$.An orthonormal set (U) has the property that each $u_i \in U$ has a norm of 1. Then for any f the orthogonal projection of f on U is

$$\widehat{f} = \sum \alpha_i u_i,$$

where

$$\alpha_i = \langle u_i, f \rangle .$$

If f belongs to the subspace generated by U, then $f = \widehat{f}$. Otherwise $\widehat{f}$ is the closest point in the subspace to f.

An important fact is that the set

$$1, \cos(2\pi kx), \sin(2\pi kx), k = \pm 1, \pm 2$$

is an orthogonal basis of $L_2(a, b)$, the space of square integrable functions on the interval (a, b). This is the foundation of Fourier-analysis.

For mathematical reasons even for studying real functions one considers frequently complex L_2 spaces where

$$\exp(2\pi ikt), k = 0, \pm 1, \pm 2, ...$$

is an orthogonal basis.

An important theorem states that if $f(t)$ square integrable and periodic with period T then there exists the Fourier-series representation:

$$
\begin{aligned}
f(t) &= \sum_{k=-\infty}^{\infty} c_k \exp(2\pi i k t/T), \\
c_k &= \frac{1}{2\pi} \int_{-T}^{T} f(t) \exp(-2\pi i k t/T) dt.
\end{aligned}
$$

Equivalently if (a, b) a finite interval and $f(t)$ is square integrable on (a, b) the same statement is essentially true.

This is in fact a projection on a specific orthonormal basis. Via a limiting argument one can obtain another fundamental result: for $L_2(-\infty, \infty)$ there exists the Fourier-transform:

$$
\widehat{f}(\omega) = \int_{-\infty}^{\infty} f(t) \exp(-2\pi i \omega t) dt
$$

and the inverse transform:

$$
f(t) = \int_{-\pi}^{\pi} \widehat{f}(\omega) \exp(2\pi i \omega t) d\omega,
$$

In words it means that for any function of "time" there eists an equivalent representation in "frequencies". The Fourier transform has huge significance theoretically, and also in practical "analogue" signal processing.

In statistical practice with a finite signal one can define the discrete Fourier transform (DFT) as

$$
\widehat{f}(k) = \sum_{t=1}^{T} f(t) \exp(-2\pi i k t), k = 0, 1, ...T - 1
$$

and its inverse transform as

$$
f(t) = \sum_{k=0}^{T-1} \widehat{f}(k) \exp(2\pi i k t).
$$

These can be regarded as approximations to the Fourier and inverse-Fourier transforms, respectively.

As one can see the parameters can be derived from a "perfect regression" :

$$
f_t = a_0 + \sum_{k=1}^{(T-1)/2} (a_k \cos(2\pi t \frac{k}{T}) + b_k \sin(2\pi t \frac{k}{T})),
$$

as because of orthogonality:

$$
\begin{aligned}
a_k &= \frac{T}{2}\sum_{t=1}^{T} x_t \cos(2\pi t\frac{k}{T}) \\
b_k &= \frac{T}{2}\sum_{t=1}^{T} x_t \sin(2\pi t\frac{k}{T}).
\end{aligned}
$$

Then $\widehat{f}(k) = a_k + ib_k$.

6.6.3 Fourier-analysis and time series analysis

It is easy to see that no trajectory of a stationary process can have a Fourier-transform. The autocorrelation function has one, however, if it is absolute summable:

$$
\begin{aligned}
f(\omega) &= \sum_{k=-\infty}^{\infty} \gamma_k \exp(-2\pi ik\omega) \\
\gamma_k &= \int f(\omega)\exp(2\pi ik\omega)d\omega, k = 0, \pm 1...
\end{aligned}
$$

This is called the spectrum or spectral density of the corresponding process. It is true that

$$
\begin{aligned}
f(\omega) &= f(-\omega) \\
\gamma_0 &= \int_{-0.5}^{0.5} f(\omega)d\omega.
\end{aligned}
$$

One can prove that frequencies outside $(-0.5, 0.5)$ can be neglected, because of periodicity. We can conclude that the spectrum and the autocovariance function contain the same information in the absolute summable case..

The fundamental Spectral Representation Theorem is the frequency domain equivalent of the Wold Representation Theorem. It asserts that any covariance-stationary process can be represented as a stochastic integral

$$
y_t = \mu + \int_0^{\pi} \left[A(\omega)\cos(\omega t) + B(\omega)\sin(\omega t)\right] d\omega,
$$

where and $A(\omega)$ and $B(\omega)$ are continuous "time" stochastic processes with certain properties. (The definition of a stochastic integral is outside the scope of these lecture notes.)

6.6.4 Statistical problem: how to estimate the spectrum?

One approach is parametric estimation. For instance, we may assume that the process in question is $ARMA(p,q)$, then we derive the theoretical spectrum which provides a correspondence to the ARMA parameters. After estimating the ARMA parameters in some traditional fashion we plug the estimates into the theoretical spectrum.

Another popular method is a non-parametric estimate.

The DFT of the sample as defined above looks like

$$\begin{aligned} d(\frac{k}{T}) &= \frac{1}{\sqrt{T}}(\sum_{t=1}^{T} x_t \cos(2\pi t\frac{k}{T}) - i\sum_{t=1}^{T} x_t \sin(2\pi t\frac{k}{T})) \\ 0 &\leq k \leq T-1. \end{aligned}$$

It can be written in matrix form as

$$\mathbf{d} = \mathbf{F}\mathbf{x},$$

where, for instance

$$F = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & \exp(-i\frac{1}{T}) & \exp(-i\frac{2}{T}) & \exp(-i\frac{3}{T}) \\ 1 & \exp(-i\frac{2}{T}) & \exp(-i\frac{4}{T}) & \exp(-i\frac{6}{T}) \\ 1 & \exp(-i\frac{3}{T}) & \exp(-i\frac{6}{T}) & \exp(-i\frac{9}{T}) \end{bmatrix},$$

Then

$$\mathbf{x} = \frac{1}{T}\overline{F}^{*}\mathbf{d},$$

as the inverse DFT "reconstructs" the orginal series. The squared DFT is called the scaled periodogram:

$$P(\frac{k}{T}) = a_k^2 + b_k^2.$$

where a_k and b_k are the real and imaginary parts of d_k. Alternatively the periodogram is the DFT of the empirical autocorrelation function.

When x_t is stationary the periodogram is stationary, too, and confidence intervals can be calculated. The periodogram is an unbiased but not consistent estimate of the spectrum. It can be made consistent via smoothing.

Spectral analysis is not so frequently used by econometricians as the the time domain methodologies. The main reason perhaps is that in contrast to many natural phenomena economic time series do not exhibit sharp differences between frequencies, in other words do not seem to exist clear cycles. However, theoreticians use frequency domain methods because certain operations are more easy to carry out in the frequency domain, and, also, because constructing and studying frequency domain filters appears to be a useful data preprocessing

strategy. For instance, the renowned band-pass filter filters out short and long frequency components from a macroeconomic time series in order to separate the "business cycle" component. Frequency domain methods are theoretically sound only if stationarity can be supposed. As in many areas this assumption is more than doubtful new methodologies have become fashionable promising "automatic" analysis for non-stationary series.

6.7 Wavelets

Signal processing applications of the Fourier-transform encountered several difficulties. Fourier-analysis is unable to detect non-stationarity, and its global nature prevents it from giving information about the local (in time) behaviour of the series. One proposed solution was the windowed Fourier transform that essentially is a slicing up in time if the window is simple. The formula is

$$\widehat{f}(\omega,\tau) = \int_{-\infty}^{\infty} f(t)w(t-\tau)\exp(-2\pi i\omega t)dt,$$

where $w(t-\tau)$ is a "window" function. However it is not self-adapting, one has to discover the appropriate shape of the window function. The wavelet transform can enable us to rectify these problems at the cost of increased freedom of choice: whereas the Fouruer transform is essentially unique there are an infinite number of substantially different wavelet transforms.

6.7.1 The wavelet transform

The wavelet transform provides us with a decomposition of a time series into scale and time components, while the Fourier-transform gives only frequency decomposition, and the Wold Representation Theorem only time decomposition. As a simplification one could say that the wavelet transform expresses how much a time series changed around a certain date at different scales. It has been likened to a prism through which one can observe the properties of an object (the time series in our case) otherwise obscured. It is customary to relate it to the Fourier-transform that assumes a similar task, but relies on the assumption of homogeneity (stationarity), and does not account for local (localized in time) changes. In the role of prism wavelets have been proved to improve on Fourier-analysis, at least in the life and earth sciences. In other words, to characterize complex and non-stationary systems this methodology has advantages.

6.7.2 Continuous wavelet transform

It starts from the windowed Fourier transform, but replaces $w(t-\tau)\exp(-2\pi i\omega t)$ with some (time dependent) filter:

$$W(s,\tau) = \int_{-\infty}^{\infty} x(t)\frac{1}{\sqrt{s}}\psi^*(\frac{t-\tau}{s})dt,$$

where $\psi(t)$ is called a wavelet n having (somewhat simplified) properties $\int_{-\infty}^{\infty} \psi(t)dt = 0$ and $\int_{-\infty}^{\infty} |\psi(t)|^2 dt = 1$. Here the frequency of the Fourier transform is replaced with scale (s).The wavelet transform is a convolution:

$$(f * g)(\tau) = \int f(t)g(\tau - t)dt,$$

for any scales.

Continuous wavelets are highly redundant transformations, when calculated from an actual time series the computation produces a matrix with much more entries than the original series. They must be distinguished from discrete wavelets that specifically strive for data compression and are used much less in research than in engineering. In economic applications the most commonly used mother wavelet is the Morlet wavelet.

What kind of statistics can we derive from the wavelet transfom to analyze data? The Wavelet Power Spectrum (WPS) is the squared wavelet transform. WPS figures can be created with the following interpretation: a point with abscissa (time period), and ordinate (scale) expresses the power attributable to that time and scale. The integral of the WPS equals the variance of the time series, thus the WPS can be interpreted as producing variance decomposition.

If we have two series the cross wavelet transform is defined as the conjugate product of the two individual transforms. From this one can define the Wavelet Coherency (WC) measure which is similar to the cross-autocovariance function, but having also a time dimension. The cross wavelet transform makes possible the calculation of phase differences, establishing lead-lag relationships between the series. We can calculate the most powerful time and the most powerful scale statistics, where the WC values are averaged time- and scale-wise, and then arg-maxed according to time and scale, respectively.

6.7.3 The orthogonal wavelet transform

It starts from the Fourier-series, and looks for an orthogonal representation of $x(t)$:

$$x(t) = \sum_{j,k=-\infty}^{\infty} W(j,k)\varphi(t,k,j),$$

where $\{\varphi(t,k,j)\}$ is an orthonormal set of functions. It is the inverse of the wavelet transform resulting in the coefficients $W(j,k)$.

The construction of this orhogonal representation starts with a mother wavelet and a father wavelet. The simplest is the Haar mother wavelet

$$\psi(t) = \left\{ \begin{array}{c} -1, 0 \leq t < \frac{1}{2} \\ 1, \frac{1}{2} \leq t < 1 \\ 0, else \end{array} \right\}$$

and the Haar father wavelet:

$$\phi(t) = \left\{ \begin{array}{c} 1, 0 \leq t < 1 \\ 0, else \end{array} \right\}$$

The daughters are defined as

$$\psi_{j,k}(t) = 2^{\frac{j}{2}} \psi(2^j t - k),$$

whereas the scaling functions as:

$$\phi_k(t) = \phi(t - k).$$

Then the union $\{\phi_k(t)\} \cup \{\psi_{j,k}(t)\}$ for all integers j, k forms an orthonormal basis of L_2.

For practical purposes it is important that it can be proved that there is a general method for finding an orthonormal basis, starting from an appropriate father or mother wavelet.

The discrete orthogonal wavelet transform in practice For finite data we have to choose a finite basis. We assume that $n = 2^J$ and make the restriction $k = 2^j$.

For instance the 8-sample Haar-wavelet (without normalizing constants) looks like:

$$\begin{bmatrix} -1 & 1 & & & & & & \\ & & -1 & 1 & & & & \\ & & & & -1 & 1 & & \\ & & & & & & -1 & 1 \\ -1 & -1 & 1 & 1 & & & & \\ & & & & -1 & -1 & 1 & \\ -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

The first wavelet in this case is

$$\psi_{1,1}(t) = \left\{ \begin{array}{c} -1, 0 \leq t < \frac{1}{2^3} \\ 1, \frac{1}{2^3} \leq t < \frac{1}{2^2} \\ 0, else \end{array} \right\},$$

and the fifth:

$$\psi_{2,1}(t) = \left\{ \begin{array}{c} -1, 0 \leq t < \frac{1}{2^2} \\ 1, \frac{1}{2^2} \leq t < \frac{1}{2} \\ 0, else \end{array} \right\}.$$

The wavelet transform (without normalizing constants) is

$$
\begin{aligned}
&x_2 - x_1 \\
&x_4 - x_3 \\
&x_6 - x_5 \\
&x_8 - x_7 \\
&(x_3 + x_4) - (x_3 + x_4) \\
&(x_7 + x_8) - (x_5 + x_6) \\
&(x_5 + x_6 + x_7 + x_8) - (x_1 + x_2 + x_3 + x_4) \\
&(x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8).
\end{aligned}
$$

The orthogonal wavelet transform enables multiresolution analysis:

$$x = w_0 + w_1 + ... + w_J + v_J,$$

where each w_j is in the orthogonal complement of V_j (where $v_j \in V_j$.)

The wavelet transform and its inverse can be written in matrix form as

$$
\begin{aligned}
x^w &= Wx \\
x &= W^T x^w = \sum (W^T)_i x_i^w,
\end{aligned}
$$

providing a scale-wise decomposition:

$$x = \sum_j x^j.$$

This is the multiresolution analysis in practice. Its main use is data compression in computer science, but it is also used for data de-noising.

A problem with the orthogonal wavelet transform is that it is sensitive to "initial" conditions, and it requires "decimated" data. A possible solution is the Maximal Overlap DWT (MODWT). It does not restrict observations to $k = 2^j$, however it is not orthogonal and is redundant. Its construction requires artificial data. But reconstruction is possible, and gives a multiresolution analysis with preserving variance. Both the orthogonal wavelet transform and the MODWT have been used in economics for estimating regressions at different scales.

6.8 Literature

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: forecasting and control. John Wiley & Sons.

Hamilton, James Douglas. Time series analysis. Vol. 2. Princeton, NJ: Princeton university press, 1994.

Kirchgässner, Gebhard, Jürgen Wolters, and Uwe Hassler. Introduction to modern time series analysis. Springer Science & Business Media, 2012.

Percival, D. B., & Walden, A. T. (2000). Wavelet methods for time series analysis (Vol. 4). Cambridge university press.

Shumway, Robert H., and David S. Stoffer. "Time series regression and exploratory data analysis." Time series analysis and its applications. Springer New York, 2011. 47-82.