Problem Set 1: Identification & Stationary Time Series

Identification

Checking Identifiability

Exercise 1 (2 Marks)**.** Let Y, X and U be random variables where the unobservable U comes from a standard Normal distribution, i.e. $U \sim N(0,1)$, where

$$Y = \alpha + \beta X + U \tag{1}$$

Suppose we know the distribution of X (it is independent of α and β) and we know that $X \perp\!\!\!\perp U$. Are α and β identified by (1)? If yes, then prove it. If no, then display two or more values of the parameters for which the distribution of Y is the same.

Conditional Prediction

Exercise 2 (Optional)**.** A researcher observes the scores of the population of foreign Ph.D. students who take and pass the composition part of the TOEFL examination. For each student, the researcher observes the following:

$$\begin{aligned} y &= \text{the TOEFL composition score (the passing scores are 4, 5, 6)} \\ x &= \begin{cases} 1 & \text{if student has a mathematics or science bachelor's degree,} \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The population distribution $P(y,x)$ is shown below.

	Test Score			
Degree	$y=4$	$y=5$	$y=6$	Totals
$x=0$	0.20	0.40	0.15	0.75
$x=1$	0.05	0.10	0.10	0.25
Total	0.25	0.50	0.25	1.00

a. Find a best predictor of y conditional on $x=0$, under absolute loss.

b. Find a best predictor of y conditional on $x=1$, under square loss.

c. Find a best predictor of x conditional on $y=$ (4 or 5), under square loss.

d. Find a best predictor of x conditional on $y=6$, under absolute loss.

e. Observing that $P(Y=6|X=0)=0.2$ and $P(y=6|X=1)=0.4$, a researcher states the following:

> The data indicate that receiving a mathematics or science bachelor's degree substantially increases the chance that a student obtains the highest test score. The estimated effect of a math/science degree is to increase the probability of scoring 6 from 0.2 to 0.4.

Does this statement accurately describe the empirical finding? Explain.

Exercise 3 (5 Marks)**.** An election official wants to make a point prediction of the number of persons in a tiny village who will vote in an election. The village has two eligible voters, denoted $j=1$ and 2. Let $y_j=1$ if person j will vote and $y_j=0$ otherwise. The official knows that the voting probabilities for the two voters are

$$P(y_1=1)=P(y_2=1)=0.5.$$

a. Assume that $P(y_1, y_2) = P(y_1)P(y_2)$. Find a best predictor of $y_1 + y_2$ under square loss. Under absolute loss. (3 Marks)

b. Assume instead that $P(Y_1 = Y_2) = 1$. Now find a best predictor of $Y_1 + Y_2$ under square loss. Under absolute loss. (2 Marks)

Exercise 4 (5 Marks)**.** `INPUTM12.txt` is a data file that contains 869 observations of American white male respondents in the National Longitudinal Study of Youth (NLSY). Each record consists of values for the variables (y, z, f, m), which are defined by:

$$\begin{aligned} y &= \text{indicator of high school completion (1 = yes, 0 = no)} \\ z &= \text{indicator of family status at age 14 (1 = intact, 0 = non-intact family)} \\ f &= \text{father's years of schooling} \\ m &= \text{mother's years of schooling} \end{aligned}$$

Suppose that the mother of an American white male has 12 years of schooling and you are asked to predict high school graduation. Assume that the 869 observations are a random sample of American white males. Use MATLAB software to do the following:

1. Estimate the best linear predictor of y given ($m = 12$) under square loss, by ordinary least squares. (1 Mark)
2. Compute kernel estimates of $E(y|m = 12)$ using uniform and Gaussian kernels and bandwidths 0.5, 1.5 and 4.5; hence, there are six estimates in total. (3 Marks)
3. Discuss the estimates computed under 1 and 2. (1 Mark)

Incomplete Data & Stochastic Dominance

Exercise 5 (3 Marks)**.** Consider the example of the wage reservation model. That is, contemplate $P(y, z, x, R)$ where R denotes reservation wage, x are covariates, y is wage (sometimes observed and sometimes unobserved) and z is defined by

$$z = \begin{cases} 1 & \text{if } y > R \\ 0 & \text{if } y < R \\ \in \{0, 1\} & \text{if } y = R \end{cases}$$

1. Express the identification region for $P(y|x)$ assuming we know $P(y > R|x)$, $P(y < R|x)$ and $P(y|x, y > R)$. (1 Mark)
2. Now assume we have a homogeneous reservation wage, i.e. suppose for a given x, the reservation wage is the smallest observed wage $y^*(x)$, i.e. R is the same for all people. Show that $P(y \leq t|x)$ is point identified when $t > y^*(x)$. (1 Mark)
3. Show that in this case $P(y \leq t|x)$ under the assumption of missingness at random stochastically dominates $P(y \leq t|x)$ under the assumption of homogeneous reservation wage. (1 Mark)

More Distributional Assumptions

Exercise 6 (3 Marks)**.** Parametric assumptions are weaker than distributional assumptions, so they may be more credible. In this question, we will look specifically at what is to be added when we assume the assumption of means missing monotonically. Recall from lectures that the weakening of means missing at random (equality) to means missing monotonically (inequality) gives

$$E[g(y)|x, w, z = 1] \geq E[g(y)|x, w, z = 0]$$

The sign of the inequality could be reversed. For example, let $g(y) = y$ and consider inference on a wage regression. This assumption could mean that the mean market wage of those that work is no less than the mean market wage of those that do not work. We need a context to interpret this.
Assume $g_0 \leq E[g(Y)] \leq g_1$. Compare the identification region for $E[g(y)]$ without any assumptions using the data alone to that which you obtain combining the data with the assumption.
Optional: compare the identification region for $E[g(y)]$ without any assumptions using the data alone to that which you obtain using the assumption of means missing at random and also to that which you obtain using the assumption of mean independence.

Decomposition of Mixtures

Exercise 7 (4 Marks)**.** One specific problem of political science and sociology is the ecological inference problem. Let us look at the analysis of voting behaviour and suppose we are interested in figuring out the voting behaviour of minorities. Let y denote the voting behaviour on some election and w be personal covariates. For the purposes of this question, let

$$y = \begin{cases} 1 & \text{democrat} \\ 0 & \text{republican} \end{cases}$$

$$w = \begin{cases} 1 & \text{white} \\ 0 & \text{black} \end{cases}$$

Assume there are no other parties and that everyone votes. We may get $P(y = 1)$ from election records and $P(w = 1)$ from the census. We do not have data on $P(y|w)$. Duncan & Davis (1953) solved a similar problem in partial identification, where the motivation was a lack of surveys.

1. Use the law of total probability to expand $P(y = 1)$ and identify what quantities we know and what quantities we do not know. (1 Mark)
2. Let $P(w = 0) = p$ and express $P(y = 1|w = 1)$ in terms of $P(y = 1)$, $P(y = 1|w = 0)$ and p. (1 Mark)
3. Write down the identification region for $P(y = 1|w = 1)$. When will this be uninformative? (2 Marks)

Treatment Response with External Validity

Exercise 8 (6 Marks)**.** Consider the problem of how sentencing juvenile offenders may affect their future criminality. Suppose we have available data on the sentencing and recidivism of males in Ireland who were born from 1980 through 1985 and who were convicted of offenses before they reached age 16. Let $t = b$ denote confinement in residential facilities and $t = a$ denote sentences that do not involve residential confinement. The outcome of interest is y defined by:

$$y = \begin{cases} 1 & \text{offender is not convicted of a subsequent crime the in five-year period following sentencing} \\ 0 & \text{offender is convicted of a subsequent crime in the five-year period following sentencing} \end{cases}$$

We have data for the study population as follows:

$$\begin{aligned} P(t = b) &= 0.10 \\ P(y = 0) &= 0.65 \\ P(y = 0|t = b) &= 0.75 \\ P(y = 0|t = a) &= 0.6 \end{aligned}$$

Consider two alternative policies: one mandating residential treatment for all offenders and the other mandating nonresidential treatment. The recidivism probabilities under these policies are $P[y(b) = 0]$ and $P[y(a) = 0]$, respectively.

1. If you assumed that judges in Ireland either purposefully or effectively sentence offenders at random to residential and nonresidential treatments, what could you conclude regarding $P[y(b) = 0]$ and $P[y(a) = 0]$? (1 Mark)
2. What would be the identification regions for these potential recidivism probabilities using the empirical evidence alone? (1 Mark)
3. What are the widths of the two intervals you calculated in 2? If they differ, why do they differ? If they do not differ, why do they not differ? (1 Mark)
4. The average treatment effect in this setting is the difference in recidivism probabilities under the two alternative sentencing policies, i.e. $P[y(b) = 0] - P[y(a) = 0]$. Calculate the identification region for average treatment effect using the data alone. What is the width of this interval? Does it contain zero? Explain. (1 Mark)
5. Calculate the average treatment effect in 2 under the assumption of treatment at random, i.e. under $P[y(a)|t = a] = P[y(a)|t = b]$ and $P[y(b)|t = a] = P[y(b)|t = b]$. (1 Mark)
6. Finally, suppose that a legal researcher wants to use this data to support the abolition of sentences confining juvenile offenders to residences. In particular, she states the following:

   > Data indicate that juvenile offenders who are not sentenced to residential confinement have a lower probability of committing future crimes. The effect of nonresidential treatment is to lower the probability of juvenile offenders committing future crimes from 0.77 to 0.59.

   Does this statement accurately describe the empirical findings? Explain. (1 Mark)

Monotone Treatment Response & Monotone Treatment Selection

Exercise 9 (2 Marks)**.** Consider the returns to education. Let $y(t)$ be the wage response to t years of schooling and assume the shape restriction of monotone treatment response (MTR), i.e.

$$t \geq s \Longrightarrow y_j(t) \geq y_j(s)$$

Let $y \in [y_0, y_1]$ and z denote received treatment. Further suppose that we take the logarithm of the wage, $f(y(t))$ so $f : Y \longrightarrow \mathbb{R}$ is a weakly increasing function. Note that $E[f(y(t))]$ respects stochastic dominance.

1. Compute the identification region for $E[f(y(t))]$ without any assumptions and with the assumption of MTR. (1 Mark)
2. MTR is nonrefutable and it enables partial prediction of outcomes for proposed new treatments that have never been used in practice. It is a lot weaker of an assumption than traditional econometric restrictions of linearity. A related assumption is that of monotone treatment selection (MTS):

$$s' \geq s \Longrightarrow E[y(t)|z = s'] \geq E[y(t)|z = s]$$

	Treatment		
years of life after treatment	$(Z = a)$	$(Z = b)$	total
$Y = 0$	.10	.12	.22
$Y = 1$	.25	.30	.55
$Y = 2$ to 10	.15	.08	.23
total	.50	.50	1

Table 1: Treatment under ambiguity.

Now suppose that instead of $E[f(y(t))]$, we are interested in $E[y(t)]$. Derive the identification region for $E[y(t)]$ under MTS. (1 Mark)

Planning Under Ambiguity

Exercise 10 (5 Marks)**.** There are two treatments for patients diagnosed with a disease $t = a$ and $t = b$. The patients in a study population have been treated with $Z = a$ for half of the patients and $Z = b$ for the remaining half. A physician obtains data on these treatment decisions and observes partial data on the number of years, denoted Y, that each patient lives after treatment. Table 1 shows the available data on the distribution of different values of Y.

1. Optional: Given the available data, what can the physician deduce about the average treatment effect,

$$E[Y(b)] - E[Y(a)]$$

Note: for the rest of the exercise, use the bounds you would get from this part:

$$E[Y(b)] \in [0.46, 6.1]$$
$$E[Y(a)] \in [0.55, 6.75]$$

2. What is the maximin treatment rule? (1 Mark)
3. What is the minimax-regret treatment rule? (1 Mark)
4. Suppose that the physician declares himself to be a Bayesian and chooses to assign all new patients to treatment b. What can you conclude about his subjective beliefs regarding relevant unobserved quantiles? Be specific. (3 Marks)

MATLAB

Classical Linear Regression

Exercise 11 (20 Marks)**.** Consider the Classical Linear Regression model in matrix form,

$$\underbrace{y}_{Tx1} = \underbrace{x}_{TxK}\underbrace{\beta}_{Kx1} + \underbrace{e}_{Tx1} \tag{2}$$

(a) Describe the main assumptions of the CLRM and specify the variance-covariance structure of the disturbances e. (1 Mark)

(b) Derive the OLS estimator $\hat{\beta}$ and show that $\hat{\beta}$ is unbiased. (1 Mark)

(c) Assume that $e \sim N(0, 0.7)$, $\beta = 2.0$ and $x = (1, ..., 1)?, T = 500$. Write a program in Matlab to show that $E[\hat{\beta}] = \beta$ and plot the density of the estimated betas. (4 Marks)

(d) Derive the analytical covariance of $\hat{\beta}$ denoted $\sum_{\hat{\beta}}$. (1 Mark)

(e) Summarize and explain the main properties of $\hat{\beta}$. (1 Mark)

(f) Derive an unbiased estimator for $Var[e] = \sigma^2$ denoted $\hat{\sigma}2$ and show that $E[\sigma^2] = \sigma^2$. (1 Mark)

(g) Derive the R^2 of the CLRM and explain the intuition behind it. Why do weneed to adjust the R^2 of a regression model? (2 Marks)

(h) Derive the likelihood function for estimating the parameters β and σ^2 of the CLRM via ML and explain the intuition behind Maximum Likelihood (ML) estimation. (2 Marks)

(i) Derive the ML estimators $\tilde{\beta}$ and $\tilde{\sigma}^2$ and show that $E[\tilde{\beta}] = \beta$ but that $\tilde{\sigma}^2$ is a biased estimator for σ^2. (1 Mark)

(j) Using the same values as in (c), write a program in Matlab to show that $E[\tilde{\beta}] = \beta$ and plot the density of the estimated betas. (4 Marks)

(k) Summarize and explain the main properties of ML estimation. (1 Mark)

(l) Show that the ratio $t = (\tilde{\beta}_k - \beta)/\hat{\sigma}_{\hat{\beta}}$ is t-distributed with $(T - K)$ degrees of freedom. (1 Mark)

ARMA Models

Exercise 12 (40 Marks)**.** For this exercise you will need the dataset tsdata.mat and the problems MUST be implemented in Matlab where indicated. For this you will need to provide your Matlab program in a separate sheet and please highlight the changes you did to the original program. Since the following exercises should be implemented for three different countries, you only need to provide the Matlab code for one country but the necessary output should be provided for each country. Let the stock market indices be denoted as P_{1t} for the US, P_{2t} for Germany, P_{3t} for the UK and similarly for the dividend yields as DP_{1t}, DP_{2t} and DP_{3t}. Construct the dividend series for each country as $D_{it} = P_{it}(DP_{it}/100)$. Construct log return series, dividend growth series and log dividend yield series for each country as $\triangle p_{it} = lnP_{it} - lnP_{it-1}, \triangle d_{it} = lnD_{it} - lnD_{it-1}$.

(a) Consider the following AR(2) model of log returns for each of the countries:

$$\triangle p_{it} = \phi_{0i} + \phi_{1i} p_{it-1} + \phi_{2i} \triangle p_{it-2} + e_{it}, e_{it} \sim (0, \sigma_i^2). \tag{3}$$

Estimate the parameter vector $\phi_i = (\phi_{0i}, \phi_{1i}, \phi_{2i})'$ for countries $i = 1, 2, 3$ via OLS in Matlab. Compute the corresponding t-ratios, R^2, adjusted R^2 and information criteria of the model. Diagnose the estimated residuals e_{it} for autocorrelation, normality, conditional heteroskedasticity and misspecification. According to your results are stock returns predictable from past returns in any of the countries? Is the AR(2) model above more/less appropriate than an AR(1) model in the countries considered? Justify your answers. (10 Marks)

(b) Consider the following MA(2) dividend growth model for each of the countries:

$$\triangle d_{it} = \delta_{0i} + \delta_{1i} e_{it-1} + \delta_{2i} e_{it-2} + e_{it}, e_{it} \sim iidN(0, \sigma_i^2). \tag{4}$$

Estimate the parameter vector $\delta_i = (\delta_{0i}, \delta_{1i}, \delta_{2i})'$ for countries $i = 1, 2, 3$ via Maximum Likelihood in Matlab. Compute the corresponding t-ratios, R^2, adjusted R^2 and information criteria of the model. Diagnose the estimated residuals $\hat{e}_{it}$ for autocorrelation, normality, conditional heteroskedasticity and misspecification. According to your results is dividend growth predictable from past dividend growth innovations in any of the countries? Is the MA(2) model above more/less appropriate than an MA(1) model in the countries considered? Justify your answers. (10 Marks)

(c) Consider the following ARMA(1,1) stock return model for each of the countries:

$$\triangle p_{it} = \phi_{0i} + \phi_{1i}\triangle p_{it-1} + \delta_{1i}e_{it-1} + e_{it}, e_{it} \sim iidN(0, \sigma_i^2). \tag{5}$$

Estimate the parameter vector $\phi_i = (\phi_{0i}, \phi_{1i}, \delta_{1i})'$ for $i = 1, 2, 3$ in Matlab. Compute the corresponding t-ratios, R^2, adjusted R^2 and information criteria of the model. Diagnose the estimated residuals $\hat{e}_{it}$ for autocorrelation, normality, conditional heteroskedasticity and misspecification. According to your results, would you choose the ARMA(1,1) or the ARMA(2,2) in practice to model asset returns in the 3 countries considered? Explain your answer. (10 Marks)

(d) Consider the AR(2) model,

$$y_t = \phi_0 + \phi_1 yt - 1 + \phi_2 y_{t-2} + e_t, e_t \sim (0, \sigma_e^2). \tag{6}$$

Derive the the analytical unconditional mean ($\mu = E[y_t]$), unconditional variance ($\gamma_0 = Var[y_t]$) and autocorrelation function ($\rho_h = Corr(y_t, y_{t-h})$) of the above model. Simulate the above model in Matlab with $\phi = (\phi_0, \phi_1, \phi_2)' = (0.1, 0.8, 0.1)', e_t \sim N(0, 0.85)$ and T = 500 and plot the simulated series, autocorrelation function and partial autocorrelation function of the simulated series. Explain your results. (10 Marks)

Stationary Time Series

ADL models

Exercise 13 (3 Marks)**.** In the autoregressive distributed lag model

$$y_t = 0.9y_{t-1} - 0.2y_{t-2} + 3x_{t-1} + u_t$$

where u_t is a zero mean stationary distrurbance term, find

(a) the total multiplier (1 Mark)

(b) the mean lag (1 Mark)

(c) the coefficients of x_{t-j} for $j = 0, 1, 2$ (1 Mark)

Exercise 14 (Optional)**.** The distributed lag regression model

$$y_t = \delta_0 x_t + \delta_1 x_{t-1} + \delta_2 x_{t-2} + \epsilon_t$$

can be re-written as

$$y_t = \delta_0 * \Delta x_t + \delta_j * \Delta x_{t-1} + \delta_2 * x_{t-2} + \epsilon_t$$

Express the new parameters in terms of the original parameters and explain how they may be interpreted. Are there any practical advantages to working with the re-parameterised model?

Forecasting

MA Models

Exercise 15 (2 Marks)**.** If $\epsilon_T = 1.2$ make predictions 1 and 2 steps ahead from the model

$$y_t = 5 + \epsilon_t + 0.5\epsilon_{t-1} \quad t = 1, \ldots, T$$

What is the prediction MSE?

ARMA Models

Exercise 16 (3 Marks)**.** Given that $y_T = 2.0$, $y_{T-1} = 1.0$, and $\epsilon_T = 0.5$, make predictions 1, 2 and 3 steps ahead from the model

$$y_t = 0.6y_{t-1} + 0.2y_{t-2} + \epsilon_t + 0.6\epsilon_{t-1} \quad t = 1, \ldots, T$$

Minimum MSE Forecasts

Exercise 17 (2 Marks)**.** Exerise 1 from chapter 5 of the notes: prove that a quadratic loss function implies that associated risk will be the mean square error. Furthermore, prove that under a quadratic loss function the mean is the minimum mean square error forecast.

Forecasting: Estimation, Assessment & Using Many Predictors

Exercise 18 (15 Marks)**.**

1. Suppose for the purposes of forecasting, you were asked to estimate parameters of a model (say an AR(1) for simplicity) that you worried was suffering from misspecification issues. How might you decide between an iterated approach and a direct approach and how do these two methods differ? Discuss some econometric issues that might arise if you were considering using real time data as opposed to historical data. (5 Marks)

2. Research economists uninterested in forecasting have nothing to gain from forecast assessment tools. Discuss. (5 Marks)

3. Using lots of variables for forecasting would violate the principle of parsimony. Is this statement necessarily correct? Explain. (5 Marks)