SOLUTION

EC7004, Michael Curran
HT 2013

TA Clemens Struck
January 30, 2013

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.

Solution 1 (Identifiability)**.**

Claim 1. $\theta = (\alpha, \beta)$ is identified by (1).

Proof. If θ is not identified, then for any θ, there exists $\theta' \neq \theta$ such that

$$P(Y \leq t|X;\theta) = P(Y \leq t|X;\theta')$$

Since we know the distribution of X, we can restrict ourselves only to conditional distributions, since we know the distribution of x.

$$\begin{aligned} P(Y \leq t|X;\theta) &= P(\alpha + \beta X + U \leq t|X;\theta) \\ &= P(U \leq t - \alpha - \beta X|X;\theta) \\ &= \Phi(t - \alpha - \beta X) \end{aligned}$$

where Φ is the cumulative Normal distribution function. Suppose that for all X

$$\Phi(t - (\alpha + \beta X)) = \Phi(t - (\alpha' + \beta' X))$$

Since Φ is one-to-one, it follows that for all X

$$\begin{aligned} & t - (\alpha + \beta X) = t - (\alpha' + \beta' X) \\ \implies & (\beta' - \beta)X = \alpha - \alpha' \\ \implies & \alpha = \alpha' \text{ and } \beta = \beta' \end{aligned}$$

□

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 $y=(4 \text{ or } 5)$, 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.

Solution 2 (Conditional Prediction (i))**.**

a. The best predictor of Y conditional on X under *absolute loss* is the *median* of Y conditional on X, where the median is defined to be

$$M(Y|X=x) = \min_t : P(Y \leq t|X=x) \geq \frac{1}{2} \tag{2}$$

We need to find the probability density function (pdf) of $(Y|X=0)$ to calculate the cumulative density function (cdf) of $(Y|X=0)$ so that we can find $M(Y|X=0)$ as defined by (2). The pdf of $(Y|X=0)$ is given by

$$P(Y=4|X=0) = \frac{P(Y=4,X=0)}{P(X=0)} = \frac{0.20}{0.75} = 0.27$$
$$P(Y=5|X=0) = \frac{P(Y=5,X=0)}{P(X=0)} = \frac{0.40}{0.75} = 0.53$$
$$P(Y=6|X=0) = \frac{P(Y=6,X=0)}{P(X=0)} = \frac{0.15}{0.75} = 0.20$$

So, the cdf of $(Y|X=0)$ is given by

$$\begin{aligned} P(Y \leq 4|X=0) &= 0.27 \\ P(Y \leq 5|X=0) &= 0.80 \\ P(Y \leq 6|X=0) &= 1.00 \end{aligned} \tag{3}$$

From the cdf, we can see that $\underline{M(Y|X=0)=5}$, which is the best predictor under absolute loss.

b. The best predictor of $(Y|X)$ under *square loss* is the *mean*, $E(Y|X)$. To calculate the best predictor of $(Y|X=1)$ under square loss, i.e. to calculate $E(Y|X=1)=\sum_y y \cdot P(Y|X=1)$, we need to find the pdf of $(Y|X=1)$. The pdf of $(Y|X=1)$ is given by

$$P(Y=4|X=1) = \frac{P(Y=4|X=1)}{P(X=1)} = \frac{0.05}{0.25} = 0.2$$
$$P(Y=5|X=1) = \frac{P(Y=5|X=1)}{P(X=1)} = \frac{0.10}{0.25} = 0.4$$
$$P(Y=6|X=1) = \frac{P(Y=6|X=1)}{P(X=1)} = \frac{0.10}{0.25} = 0.4$$

Now to calculate the mean

$$\begin{aligned} E(Y|X=1) &= \sum_y y \cdot P(Y|X=1) \\ &= 4 \times 0.2 + 5 \times 0.4 + 6 \times 0.4 \\ &= 5.2 \end{aligned} \tag{4}$$

Therefore, the best predictor of $(Y|X=1)$ under square loss is $\underline{E(Y|X=1) = 5.2}$.

c. The pdf of $(X|Y=4 \vee Y=5)$ is given by

$$\begin{aligned} P(X=0|Y=4 \vee Y=5) &= \frac{P(X=0, Y=4 \vee Y=5)}{P(Y=4 \vee Y=5)} \\ &= \frac{0.60}{0.75} = 0.8 \\ P(X=1|Y=4 \vee Y=5) &= \frac{P(X=1, Y=4 \vee Y=5)}{P(Y=4 \vee Y=5)} \\ &= \frac{0.15}{0.75} = 0.2 \end{aligned}$$

The best predictor of $(X|Y=4 \vee Y=5)$ under square loss is the mean, which is given by

$$\begin{aligned} E[X|Y=4 \vee Y=5] &= \sum_x x \cdot P(X|Y=4 \vee Y=5) \\ &= 0 \times 0.8 + 1 \times 0.2 \\ &= 0.2 \end{aligned}$$

d. The pdf of $(X|Y=6)$ is given by

$$\begin{aligned} P(X=0|Y=6) &= \frac{P(X=0, Y=6)}{P(Y=6)} = \frac{0.15}{0.25} = 0.6 \\ P(X=1|Y=6) &= \frac{P(X=1, Y=6)}{P(Y=6)} = \frac{0.10}{0.25} = 0.4 \end{aligned}$$

So, the cdf of $(X|Y=6)$ is given by

$$\begin{aligned} P(X \leq 0|y=6) &= 0.6 \\ P(X \leq 1|Y=6) &= 1 \end{aligned} \tag{5}$$

The best predictor of $(X|Y=6)$, under absolute loss is the median, which is given by

$$M(X|Y=6) = 0$$

e. No, this statement does not accurately describe the empirical finding.
We can only say that the foreign PhD students having maths or science bachelor's degrees on average scored higher than those who did not have such a background. We cannot say that the possession of a maths or science degree increased the probability of a student scoring a 6.
Asking what would happen to this $E(Y|X)$ when we vary X is akin to a hypothetical change in X, where we have no data and so the researcher has confused correlation with causation and has used a counterfactual (expressing what has not happened but what might or would happen if circumstances, i.e. data, were different). The researcher is in effect extrapolating using the assumption of external validity, which is undermined by the fact that we are only looking at *foreign PhD students* and have no data on

the rest of the population of students at large.
However, if the students were randomly assigned bachelor's degrees in maths / science (an impossibility in this case, but possible in more general cases where X could include randomly distributing answers to the test to some of the students), then the researcher would be correct in saying that an increase in that covariate (e.g. having the answers to the test prior to the test) increases the probability that a student will do better on average than a student who does not have answers to the test. But since the distribution of maths and science degrees is non-random and we are dealing with what actually happened (descriptive) we cannot say that having a maths or science degree increases the probability that a student scores a 6.

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)

Solution 3 (Conditional Prediction (ii))**.**

a. Under *square loss*, the best predictor of Y is $E(Y|X)$.

$$\begin{aligned}
E[Y_1 + Y_2] &= E(Y_1) + E(Y_2) \quad \because E \text{ linear operator} \\
\text{Now} \quad E(Y_1) &= \sum_{y_1} y_1 \cdot P(Y_1) \\
&= 0 \times P(Y_1 = 0) + 1 \times P(Y_1 = 1) \\
&= 0 \times \frac{1}{2} + 1 \times \frac{1}{2} \\
&= \frac{1}{2} \\
\text{Similarly} \quad E(Y_2) &= \frac{1}{2} \\
\therefore \quad E(Y_1 + Y_2) &= 1
\end{aligned}$$

Note that $P(y_1, y_2) = P(y_1)P(y_2)$ implies that the two events, Y_1 and Y_2 are independent. The pdf of $Y_1 + Y_2$ is given by

$Y_1 + Y_2$	$P(Y_1 + Y_2)$
0	$P(Y_1 + Y_2 = 0) = P((Y_1 = 0) \wedge (Y_2 = 0))$ $= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
1	$P(Y_1 + Y_2 = 1)$ $= P(((Y_1 = 1) \wedge (Y_2 = 0)) \vee ((Y_1 = 0) \wedge (Y_2 = 1)))$ $= \frac{1}{2} \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = \frac{1}{2}$
2	$P(Y_1 + Y_2 = 2) = P((Y_1 = 1) \wedge (Y_2 = 1))$ $= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$

$$\therefore E(Y_1 + Y_2) = 0 \times \frac{1}{4} + 1 \times \frac{1}{2} + 2 \times \frac{1}{4} = 1$$

So, the cdf of $Y_1 + Y_2$ is given by

$$\begin{aligned} P(Y_1 + Y_2 \leq 0) &= \frac{1}{4} \\ P(Y_1 + Y_2 \leq 1) &= \frac{3}{4} \\ P(Y_1 + Y_2 \leq 2) &= 1 \\ \therefore \quad M(Y_1 + Y_2) &= 1 \\ \therefore \quad E(Y_1 + Y_2) &= M(Y_1 + Y_2) \\ &= 1 \end{aligned}$$

So, the best predictor of $Y_1 + Y_2$, under square loss (the mean) is the same as the best predictor of $Y_1 + Y_2$, under absolute loss (the median): $\underline{1}$.

b. In this case, Y_1 and Y_2 can only take on some values.

$$\therefore \text{either } y_1 = y_2 = 0 \quad \vee \quad y_1 = y_2 = 1$$

$Y_1 + Y_2$	$P(Y_1 + Y_2)$
0	$\frac{1}{2}$
2	$\frac{1}{2}$

$$\begin{aligned} \therefore \quad E(Y_1 + Y_2) &= 0 \times \frac{1}{2} + 2\frac{1}{2} = 1 \\ M(Y_1 + Y_2) &= 0 \end{aligned}$$

So the best predictor of $Y_1 + Y_2$ under square loss is $\underline{1}$ and under absolute loss is $\underline{0}$.

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:

y = indicator of high school completion (1 = yes, 0 = no)
z = indicator of family status at age 14 (1 = intact, 0 = non-intact family)
f = father's years of schooling
m = mother's years of schooling

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)

Solution 4 (MATLAB – kernreg)**.**

1. The best linear predictor of $(Y|M = 12)$, under square loss is given by the mean. Using an OLS regression to estimate the coefficient of the constant, the following MATLAB code 1 produces the estimate.

```
b = regress(y,m)
yhat = b(1,1)+b(2,1)*m

plot(yhat,y)
```

Listing 1: BLP of y given $m = 12$ under square loss by OLS.

Alternatively, you could have used Stata for which the following code 2 prodcues the estimate.

```
infile y z f m using INPUTM12.txt, clear

regress y m
predict yhat
list yhat if m==12
```

Listing 2: BLP of y given $m = 12$ under square loss by OLS.

This yielded the expectation of $Y|M = 12$

$$\underline{E(Y|M = 12)} \;=\; \underline{.8434434}$$

which is the best linear predictor of $Y|M = 12$, under square loss.
Note that the OLS estimation of $E(Y|M = 12)$ reduces to the non-parametric estimator of $E(Y|M = 12)$ when the covariates have positive probability (true since M is discrete and the data contains a subset where $M = 12$):

$$\begin{aligned}
\hat{E}(Y|M = 12)_{OLS} &= \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2} \\
&= \frac{\sum_{i=1}^{N'} x_i y_i}{\sum_{i=1}^{N'} x_i^2} \\
&= \frac{\sum_{i=1}^{N} y_i \cdot 1[M_i = 12]}{\sum_{i=1}^{N} 1[M_i = 12]} \\
&= E_N(Y|M = 12)
\end{aligned}$$

where N is the sample size (869), $N' < N$ is the number of observations for which $M = 12$ and $x_i = 1\ \forall i : M_i = 12$.

2. The kernel estimates of $E(Y|M = 12)$ computed with the `kernreg` function, using the MATLAB code shown in listing 3, are recorded in table 1; alternatively, you could have used the Stata code shown in listing 4. The Uniform Kernel (local average) and Gaussian Kernel (local weighted average) estimates were calculated as

$$\theta_N(M = 12, d_N) = \frac{\sum_{i=1}^{N} y_i \cdot 1[\rho(M_i, M = 12) < d_N]}{\sum_{i=1}^{N} 1[\rho(M_i, M = 12) < d_N]} \qquad \text{(Uniform)}$$

$$E_N(Y|M = 12) = \frac{\sum_{i=1}^{N} y_i \cdot K\left[\frac{\rho(M_i, M=12)}{d_N}\right]}{\sum_{i=1}^{N} K\left[\frac{\rho(M_i, M=12)}{d_N}\right]} \qquad \text{(Gaussian)}$$

where $d_N \in \{0.5, 1.5, 4.5\}$ is the bandwidth.

```
kern(y,m,0.5,12,'rec')
kern(y,m,1.5,12,'rec')
kern(y,m,4.5,12,'rec')
kern(y,m,0.5,12,'gau')
kern(y,m,1.5,12,'gau')
kern(y,m,4.5,12,'gau')
```

Listing 3: Uniform & Gaussian kernel estimates of $E(y|m = 12)$ with different bandwidths.

```
kernreg y m, g(yubar1) at(12) w1(0.5) rec

kernreg y m, g(yubar2) at(12) w1(1.5) rec

kernreg y m, g(yubar3) at(12) w1(4.5) rec

kernreg y m, g(ygbar1) at(12) w1(0.5) gau

kernreg y m, g(ygbar2) at(12) w1(1.5) gau

kernreg y m, g(ygbar3) at(12) w1(4.5) gau

list yubar1 yubar2 yubar3 ygbar1 ygbar2 ygbar3 in 1/1
```

Listing 4: Uniform & Gaussian kernel estimates of $E(y|m = 12)$ with different bandwidths.

Table 1: Uniform & Gaussian kernel estimates of $E(Y|M = 12)$ with bandwidths: 0.5, 1.5 and 4.5

Bandwidth	Uniform Kernel	Gaussian Kernel
0.5	0.888172	0.8865797
1.5	0.8781362	0.8720608
4.5	0.8543689	0.8533347

3. The point here is to compare OLS and nonparametric estimators. Essentially, by expanding the bandwidth, we use more data and so we get closer to the OLS estimate. This would receive full credit. Extra discussion follows.
   For $d_N < 1$, since $m \in \mathbb{N}$, only i such that $m_i = m = 12$ will be given a value of one (instead of zero) by the indicator function in the estimation.
   The Gaussian Kernel will put more weight on values close to $m = 12$, so $K\left[\frac{\rho(M_i, M=12)}{d_N}\right] = K\left[\frac{\rho(M=12, M=12)}{d_N}\right]$ for i such that $m_i = m = 12$ (this will be the only i where the indicator function will be non-zero in the estimation) but unlike the Uniform Kernel estimate, the indicator function $K[\cdot]$ will take a value that will not be exactly one; thus the Gaussian Kernel estimate in the case $d_N = 0.5$ will be different to the Uniform Kernel estimate.Once the bandwidths are increased above 1, e.g. $d_N \in \{1.5, 4.5\}$, the kernel estimators allow values further from $m = 12$ to be taken into account ($m \in \{11, 12, 13\}$ for $d_N = 1.5$ and $m \in \mathbb{N} : 8 \leq m \leq 16$ for $d_N = 4.5$). In fact, irrespective of the shape of the distribution (Uniform or Gaussian), the estimates seem to be decreasing with the size of the bandwidth. This is because these kernel estimators give more weight to values further away from $M = 12$ as the bandwidth increases. Increasing the bandwidth will allow for more observations to lie within the bandwidth and so the variance will come down. However, increasing the bandwidth will lead to a bias of the estimate (the values of M no longer lying within a bandwidth that can be made arbitrarily small and still contain the values of M). The curse of dimensionality problem arises in trying to chose the bandwidth to minimise the mean square error (variance and bias) as the sample size increases, especially as the dimensions increase (our case only looks at M, so there is only one dimension, hence only d_N, rather than d_N^k.

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)

Solution 5 (Stochastic Dominance)**.**

1.

$$\begin{aligned} P(y|x) &= P(y|x, z = 1)P(z = 1|x) + P(y|x, z = 0)P(z = 0|x) \\ &= \underline{P(y|x, y > R)P(y > R|x)} + \boxed{P(y|x, y < R)}\underline{P(y < R|x)} \end{aligned}$$

where we underlined the quantities that we know. Note: usually we also draw a box around quantities we do not know – this is what you should do for the final exam.

2. When $t > y^*(x)$, $P(y \leq t|x)$ is point identified. This is because

$$P(y \leq t|x) = \underline{P(y \leq t|x, z=1)P(z=1|x)} + \underbrace{P(y \leq t|x, z=0)}_{1}\underline{P(z=0|x)}$$

$\therefore$ $P(y \leq t|x)$ is point identified.

3. Let us compare the homogeneous reservation wage assumption (HRW) with missingness at random (MAR):

$$\begin{aligned} P_M(y \leq t|x) &= P(y \leq t|x, z=1)P(z=1|x) + P(y \leq t|x, z=0)P(z=0|x) \\ &\overset{\text{MAR}}{=} P(y \leq t|x, z=1) \\ P_H(y \leq t|x) &\overset{\text{HRW}}{=} \underbrace{P(y \leq t|x, z=1)}_{\in[0,1]}P(z=1|x) + P(z=0|x) \\ P_H(y \leq t|x) &\geq P_M \end{aligned}$$

i.e. MAR stochastically dominates HRW.

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.

Solution 6 (More Assumptions)**.**
Here we assume

$$E[g(Y)|V, Z=1] \geq 1E[g(Y)|V] \geq E[g(Y)|V, Z=0]$$

For example, consider the wage regression example. Then this assumption says that those who work have a weakly higher mean market wage than those who do not work. This might be seen as a plausible assumption. This assumption is not refutable, since we do not observe $E[g(Y)|V, Z=0]$, and consequently, we do not observe $E[g(Y)|V]$ either.
Without the assumption, by the law of iterated expectations:

$$E[g(Y)] = E[g(Y)|Z=1]P(Z=1) + \underbrace{E[g(Y)|Z=0]}_{\in[g_0,g_1]}P(Z=0)$$

So, we get the identification region for $E[g(Y)]$ using data alone to be

$$H[E[g(Y)]] = [E(g(Y)|Z=1)P(Z=1) + g_0P(Z=0), E(g(Y)|Z=1)P(Z=1) + g_1P(Z=0)] \quad (6)$$

We can use the law of iterated expectations again with V to get

$$E[g(Y)|Z=i] = \sum_{v\in\mathcal{V}} E[g(Y)|Z=i, V=v]P(Z=i)$$

for $i=0,1$, where $\mathcal{V}$ is the set of values V takes. Use this result and the fact that

$$P(Z=i, V=v) = P(V=v|Z=i)P(Z=i)$$

to get that

$$E[g(Y)] = \sum_{v\in\mathcal{V}} E[g(Y)|Z=1, V=v]P(Z=1, V=v) + \underbrace{E[g(Y)|Z=0, V=v]}_{\in[g_0, g_1]} P(Z=0, V=v)$$

Observe that everything here is known except for $E[g(Y)|Z=0, V=v]$. A lower bound for this quantity is g_0, the lowest possible logical value that the mean can take. By the means missing monotonically assumption, the upper bound is $E[g(Y)|Z=1, V=v]$.
Consequentially, the identified set is

$$H_1[E(g(Y))] = \left[E(g(Y)|Z=1)P(Z=1) + g_0P(Z=0), \sum_{v\in\mathcal{V}} E[g(Y)|Z=1, V=v]P(V=v)\right]$$

The assumption means missing monotonically does not point identify $E[g(Y)]$. Compare this region to the identification region using data alone (6). The lower bound is the same, but the upper bound is strictly smaller. So, while this assumption is too weak to point-identify $E[g(Y)|X]$, it does have identifying power since $H_1 \subsetneq H$.
As for means missing at random and mean independence, see Manski (2007:69-71). Proving nonrefutability and point-identification of means missing at random and possible refutability of means independence are fair game for the final exam; again see Manski (2007:69-71).

Remark. If we assume that the distribution of $y|x$, say $P(y|x) \in \Gamma_{1Y}$, then without this assumption we know the identification region $H[P(y|x)]$ but when we add the assumption we get $H[P(y|x)] \cap \Gamma_{1Y} = H_1[P(y|x)]$, which usually produces a tighter bound for all parameters. We learn $H[P(y|x)]$ from the data. We learn Γ_{1Y} from the assumption. And we learn the intersection from combining the data with the assumption. If the intersection is a strict subset of $H[P(y|x)]$, then we can say that the assumption has 'identifying power.' If the intersection is zero, then we know we have made the wrong assumption. So, $\Gamma_{1Y} \subsetneq \Gamma_Y$ is necessary for this to be meaningful. It is important to place assumptions on $P(y|x)$ rather than on $P(y|x, z=0)$ so we can test assumptions. Remember that like with Classical hypothesis testing, the intersection may be nonzero but the hypothesis may still be false. This is another illustration of the general idea that theories are disprovable but not provable, e.g. we reject or do not reject (fail to reject) a null hypothesis – in the second case we cannot say that we accept the hypothesis. Note that the term 'nonrefutable' concerns the existence of an ex-ante probability that make the intersection zero; if yes, then we say the assumption is refutable; if no, then we say that the assumption is nonrefutable. In contrast, not refuted implies that we have actually tested the assumption using data. Finally note that 'nonrefutable' relates to the population, using deductive logic, while nontestable relates to the sample.

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)

Solution 7 (Decomposition of Mixtures)**.**

1. $$\underline{P(y=1)} \overset{\text{LTP}}{=} \boxed{P(y=1|w=1)}\underline{P(w=1)} + \boxed{P(y=1|w=0)}\underline{P(w=0)}$$

2. Letting $p = P(w = 0)$, from part 1:

$$\begin{aligned} & P(y=1) = P(y=1|w=1)(1-p) + P(y=1|w=0)p \\ \iff & P(y=1|w=1)\frac{P(y=1) - P(y=1|w=0)p}{1-p} \end{aligned}$$

3. Since $P(y = 1|w = 0) \in [0, 1]$

$$P(y=1|w=1) \in \left[\frac{P(y=1)-p}{1-p}, \frac{P(y=1)}{1-p}\right] \cap [0,1] \tag{7}$$

Note that (7) is a *sharp* bound – we can get no tighter bound from what we know. Equivalently, we could express (7) as

$$H[P(y=1|w=1)] = \left[\max\left\{\frac{P(y=1)-p}{1-p}, 0\right\}, \min\left\{\frac{P(y=1)}{1-p}, 1\right\}\right]$$

which is uninformative if p is too high: we can tell a lot about $P(y = 1|w = 0)$ but not much about $P(y = 1|w = 1)$ if $p = P(w = 0)$ is too high. Since we may safely assume $0 < p < 1$, H will be uninformative when $p \geq P(y = 1)$ and $p \geq 1 - P(y = 1)$ – this can be derived from putting the max operator on the left hand side less than 0 and the min operator on the right hand side greater than one:

$$\begin{aligned} & \frac{P(y=1)-p}{1-p} > 0 \implies P(y=1) > p \\ & \frac{P(y=1)}{1-p} < 1 \implies P(y=1) < 1-p \\ & \therefore p < P(y=1) < 1-p \end{aligned}$$

$$\begin{aligned} \therefore p &< 1-p \\ 2p &< p \\ p &< \frac{1}{2} \end{aligned}$$

Remark. This decomposition of mixtures problem essentially is an incomplete data problem where we decompose a known distribution into a known mixture of unknown distributions.

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)

Solution 8 (Treatment Response)**.**

1. Random treatment assignment (random) implies that for $t' \in \{a, b\}$:

$$P(y(t')|t = t') = P(y(t')|t \neq t')$$

So, by the law of total probability:

$$\begin{aligned} P(y(t')) &\overset{\text{LTP}}{=} P(y(t')|t = t')P(t = t') + P(y(t')|t \neq t')P(t \neq t') \\ &\overset{\text{random}}{=} \underbrace{P(y(t')|t = t')}_{P(y|t=t')}[\underbrace{P(t = t') + P(t \neq t')}_{1}] \end{aligned}$$

which holds for all $t' \in \{a, b\}$. Therefore

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

2. Without using any assumptions, again using the law of total probability:

$$\begin{aligned} P(y(b) = 0) &\overset{\text{LTP}}{=} \underbrace{P(y(b) = 0|t = b)P(t = b)}_{P(y=0|t=b)} P(t = b) + \underbrace{P(y(b) = 0|t = a)}_{\in[0,1]} \underbrace{P(t = a)}_{1-P(t=b)} \\ &= 0.75 \cdot 0.1 + [0, 1] \cdot 0.9 \\ &\in [0.075, 0.975] \\ P(y(a) = 0) &\overset{\text{LTP}}{=} \underbrace{P(y(a) = 0|t = a)}_{P(y=0|t=a)} P(t = a) + \underbrace{P(y(a) = 0|t = b)}_{\in[0,1]} P(t = b) \\ &= 0.6 \cdot 0.9 + [0, 1] \cdot 0.1 \\ &\in [0.54, 0.64] \end{aligned}$$

3. The width of $H[P(y(b) = 0)]$ is 0.9 and the width of $H[P(y(a) = 0)]$ is 0.1. In the first case, this is because this is the fraction of the study population who received treatment b and who, therefore, have unobservable outcomes under treatment b. Symmetrically, the region for $P(y(a) = 0)$ has width 0.1.

4. Let $ATE = P[y(b) = 0] - P[y(a) = 0]$. From part 2:

$$\begin{aligned} P[y(b) = 0] &\overset{2}{=} [\overbrace{0.075}^{L_b}, \overbrace{0.975}^{U_b}] \\ P[y(a) = 0] &\overset{2}{=} [\underbrace{0.54}_{L_a}, \underbrace{0.64}_{U_a}] \end{aligned}$$

The lower and upper bounds for ATE are given by

$$\begin{aligned} L_b - U_a &= -0.565 \\ U_b - L_a &= 0.435 \end{aligned}$$

$$\therefore H[ATE] = [-0.565, 0.435]$$

$H[ATE]$ necessarily has a width of one and includes zero. To explain this, note that using data alone, the hypothesis of $ATE = 0$ is not refutable. Given that counterfactual outcomes are unobserved, it is possible that $y(a) = y(b)$ for every person in the population; technically, we say that $y_j(a) = y_j(b)$ for all persons j in the population. Therefore, the hypothesis of zero average treatment effects is not refutable using empirical evidence alone. The hypothesis may be made refutable only if the evidence is combined with sufficiently strong distributional assumptions.

5. We assumed random treatment assignment in part 1 and found that $P[y(a) = 0] = 0.6$ and $P[y(b) = 0] = 0.75$ so $ATE = 0.15$, indicating that nonresidential treatment is much better than residential treatment if the objective is to minimise recidivism.

6. No, this statement does not accurately describe the empirical finding.
   We can only say that juvenile offenders who were sentenced to residential confinement had on average a higher probability of recidivism. We cannot say that the treatment of being sentenced to residential confinement increased the probability of recidivism.
   The researcher has confused correlation with causation and has used a counterfactual (expressing what has not happened but what might or would happen if circumstances, i.e. data were different). The researcher is in effect extrapolating using the assumption of external validity, which is undermined by the fact that we are only looking at juvenile offenders *in Ireland* who were born during 1980-1985. Furthermore, changing the very structure of the sentences may change how people respond; this is related to the *Lucas critique*.
   However, if the juvenile offenders were randomly sentenced to residential confinement as in part 5, then the researcher would be correct in saying that the effect of nonresidential treatment is to lower the probability of juvenile offenders committing future crimes.

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):

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

   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)

Solution 9 (MTR/MTS)**.**

1. Without any assumptions by LIE:

$$\begin{aligned} f(y_0)P(z \neq t) \; + \; & E[f(y)|z = t]P(z = t) \leq E[f(y(t))] \\ \leq \; & f(y_1)P(z \neq t) + E[f(y)|z = t]P(z = t) \\ E[f(y(t))] \overset{\text{LIE}}{=} \; & E[f(y(t))|t = z]P(t = z) + \underbrace{E[f(y(t))|t \neq z]}_{\substack{\text{not observed} \\ y \in [y_0, y_1]}} P(t \neq z) \end{aligned}$$

   With the assumption of MTR by LIE:

$$\begin{aligned} E[f(y(t))] &= E[f(y(t))|z = t]P(z = t) + E[f(y(t))|z \neq t]P(z \neq t) \\ &= E[f(y(t))|z = t]P(z = t) + E[f(y(t))|z > t]P(z > t) + E[f(y(t))|z < t]P(z < t) \end{aligned}$$

	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 2: Treatment under ambiguity.

Remember t is the treatment under consideration and $z = t$ means people are assigned treatment t. For $z > t$, these people were assigned a treatment greater than t, so the outcome lies in $[y_0, y]$. For $z < t$, these people were assigned a treatment less than t, so the outcome lies in $[y, y_1]$. So:

$$\begin{aligned}\text{Lower Bound} &= E[f(y)|z = t]P(z = t) + f(y_0)P(z > t) + E[f(y)|z < t)P(z < t) \\ &= E[f(y)|z \leq t)P(z \leq t) + f(y_0)P(z > t) \\ \text{Upper Bound} &= E[f(y)|z = t]P(z = t) + E[f(y)|z > t]P(z > t) + f(y_1)P(z < t) \\ &= E[f(y)|z \geq t]P(z \geq t) + f(y_1)P(z < t)\end{aligned}$$

2.

$$E[y(t)] \overset{\text{LIE}}{=} E[y(t)|z < t]P(z < t) + E[y(t)|z = t]P(z = t) + E[y(t)|z > t]P(z > t)$$

$$\begin{aligned}\text{Lower Bound} &= y_0 P(z > t) + E[y|z = t]P(z \leq t) \\ \text{Upper Bound} &= E[y|z = t]P(z \leq t) + y_1 P(z > t)\end{aligned}$$

Remark. For those interested, Manski & Pepper (2000) derive bounds on mean outcomes and average treatment effects under the combination of MTR & MTS. The combined assumptions have more identifying power since they are informative even if the outcome space Y is unbounded unlike the case for either assumption alone. They are also jointly refutable.

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 2 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:

$$\begin{aligned}E[Y(b)] &\in [0.46, 6.1] \\ E[Y(a)] &\in [0.55, 6.75]\end{aligned}$$

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)

Solution 10 (Treatment Under Ambiguity)**.**

1.

$$\begin{aligned} E[Y(b)] &\overset{\text{LTP}}{=} E[Y(b)|Z=a]P(Z=a) + E[Y(b)|Z=b]P(Z=b) \\ &= E[Y(b)|Z=a](.50) + E[Y|Z=b](.50) \end{aligned}$$

The only thing we know about the counterfactual quantity $E[Y(b)|Z=a]$ is that

$$E[Y(b)|Z=a] \in [0,10]$$

Since the exact value of Y is not known when $Y \geq 2$, the quantity $E[Y|Z=b]$ is not point identified. We have

$$\begin{aligned} E[Y|Z=b] = & \\ &= 0 \cdot P(Y=0|Z=b) + 1 \cdot P(Y=1|Z=b) + E[Y|Z=b, Y \in \{2,\dots,10\}] \cdot P(Y \in \{2,\dots,10\}|Z=b) \\ &= 0 + \frac{.3}{.5} + E[Y|Z=b, Y \in \{2,\dots,10\}] \cdot \frac{.08}{.50} \\ &= 0.6 + E[Y|Z=b, Y \in \{2,\dots,10\}] \cdot 0.16 \\ &\in 0.6 + [2,10] \cdot 0.16 \\ &= [0.6 + 2 \cdot 0.16, 0.6 + 10 \cdot 0.16] \\ &= [0.92, 2.2] \end{aligned}$$

Hence we have that

$$\begin{aligned} E[Y(b)] &\in [0,10](0.50) + [0.92, 2.2](0.50) \\ &= [0.46, 6.1] \end{aligned}$$

We can do an analogous calculation for $E[Y(a)]$. We have

$$\begin{aligned} E[Y(a)] &= E[Y(a)|Z=a]P(Z=a) + E[Y(a)|Z=b]P(Z=b) \\ &= E[Y|Z=a](.50) + E[Y(a)|Z=b](.50) \\ &= E[Y|Z=a](.50) + [0,10](.50) \end{aligned}$$

and

$$\begin{aligned} E[Y|Z=a] &= 0 \cdot P(Y=0|Z=a) + 1 \cdot P(Y=1|Z=a) + E[Y|Z=a, Y \in \{2,\dots,10\}] \cdot P(Y \in \{2,\dots,10\}|Z=a) \\ &= 0 + \frac{.25}{.50} + E[Y|Z=a, Y \in \{2,\dots,10\}] \cdot \frac{.15}{.50} \\ &= 0.5 + E[Y|Z=a, Y \in \{2,\dots,10\}] \cdot 0.3 \\ &\in 0.5 + [2,10] \cdot 0.3 \\ &= [0.5 + 2 \cdot 0.3, 0.5 + 10 \cdot 0.3] \\ &= [1.1, 3.5] \end{aligned}$$

Hence we have that

$$\begin{aligned} E[Y(a)] &\in [1.1, 3.5](.50) + [0,10](.50) \\ &= [0.55, 6.75] \end{aligned}$$

Putting these two regions together, we have the following bounds on the average treatment effect:

$$\begin{aligned} E[Y(b)] - E[Y(a)] &\in [.46 - 6.75, 6.1 - .55] \\ &= [-6.29, 5.55] \end{aligned}$$

2. For here and the rest of this problem, let α and β denote the true values of $E[Y(a)]$ and $E[Y(b)]$, respectively. Let $[\alpha_L, \alpha_U]$ and $[\beta_L, \beta_U]$ denote their respective identified sets. The maximin rule is

$$\delta_{\text{MM}} = \begin{cases} 0 & \text{if } \alpha_L > \beta_L \\ 1 & \text{if } \beta_L > \alpha_l \\ [0,1] & \text{if } \alpha_L = \beta_L \end{cases}$$

Since $\alpha_L = 0.55 > 0.46 = \beta_L$, choose $\delta_{\text{MM}} = 0$. Assign everyone treatment a.

3. As in Manski (2007)

$$\delta_{\text{MMR}} = \frac{\beta_U - \alpha_L}{(\alpha_U - \beta_L) + (\beta_U - \alpha_L)}$$

Using the numbers from part 1,

$$\begin{aligned} \delta_{\text{MMR}} &= \frac{6.1 - 0.55}{(6.75 - 0.46) + (6.1 - 0.55)} \\ &= \frac{5.55}{6.29 + 5.55} \\ &= 0.46875 \end{aligned}$$

The doctor will assign 46.88% of the patients to the new treatment b and the others to the status quo a.

4. Here we assume that the physician uses a Bayesian decision rule and assigns everyone to treatment b. Let π be her subjective probability distribution over the states of nature, (α, β). Her decision rule, as shown in Manski (2007), is

$$\delta_{\text{Bayes}} = \begin{cases} 1 & \text{if } E_\pi[\alpha] < E_\pi[\beta] \\ 0 & \text{if } E_\pi[\alpha] > E_\pi[\beta] \\ [0,1] & \text{if } E_\pi[\alpha] = E_\pi[\beta] \end{cases}$$

Since she assigns everyone to b, so $\delta = 1$, we know that

$$E_\pi[\beta] \geq E_\pi[\alpha]$$

Now refer back to what we did in part 1:

$$\begin{aligned} \alpha &= E[Y|Z = a](.50) + E[Y(a)|Z = b](.50) \\ \beta &= E[Y(b)|Z = a](.50) + E[Y|Z = b](.50) \end{aligned}$$

and

$$\begin{aligned} E[Y|Z = a] &= 0.5 + E[Y|Z = a, Y \in \{2, \ldots, 10\}] \cdot 0.3 \\ E[Y|Z = b] &= 0.6 + E[Y|Z = b, Y \in \{2, \ldots, 10\}] \cdot 0.16 \end{aligned}$$

Thus

$$\begin{aligned} \alpha &= (0.5 + E[Y|Z = a, Y \in \{2, \ldots, 10\}] \cdot 0.3)\,(.50) + E[Y(a)|Z = b](.50) \\ \beta &= E[Y(b)|Z = a](.50) + (0.6 + E[Y|Z = b, Y \in \{2, \ldots, 10\}] \cdot 0.16)\,(.50) \end{aligned}$$

Simplifying this,

$$\alpha = 0.25 + E[Y|Z = a, Y \in \{2, \ldots, 10\}] \cdot 0.15 + E[Y(a)|Z = b](.50)$$
$$\beta = E[Y(b)|Z = a](.50) + 0.3 + E[Y|Z = b, Y \in \{2, \ldots, 10\}] \cdot 0.08$$

Here we have expressed the quantities α and β in terms of the four quantities for which our data have nothing to say about. All we know are the logical bounds

$$E[Y|Z = a, Y \in \{2, \ldots, 10\}] \in [2, 10]$$
$$E[Y|Z = b, Y \in \{2, \ldots, 10\}] \in [2, 10]$$
$$E[Y(a)|Z = b] \in [0, 10]$$
$$E[Y(b)|Z = a] \in [0, 10]$$

Since the physician is Bayesian, she has subjective beliefs about these four quantities. Denote them as above, but with π subscripts. Since she assigns everyone to b, we know that $E_\pi[\beta] \geq E_\pi[\alpha]$, as mentioned already, which implies that

$$E_\pi[Y(b)|Z = a](.50) + 0.3 + E_\pi[Y|Z = b, Y \in \{2, \ldots, 10\}] \cdot 0.08$$
$$\geq 0.25 + E_\pi[Y|Z = a, Y \in \{2, \ldots, 10\}] \cdot 0.15 + E_\pi[Y(a)|Z = b](.50)$$

Thus, we know that her beliefs about these four means must satisfy this inequality, as well as the logical bounds mentioned above.

Remark. Notice our solution strategy. We wrote the quantities of interest, α and β, in terms of the 'fundamental' unknowns: those quantities about which the data has nothing to say. We then used the decision rule to make a conclusion about the physician's belief about these fundamental unknowns. This is what the question refers to when it asks us to 'be specific'. Here is a less specific answer which, although correct, is not as precise as we could be: π is any probability distribution on (α, β) with support $[\alpha_L, \alpha_U] \times [\beta_L \beta_U]$ and such that $E_\pi[\beta] \geq E_\pi[\alpha]$.

MATLAB

Classical Linear Regression

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

$$\underbrace{y}_{Tx1} = \underbrace{x}_{TxK} \underbrace{\beta}_{Kx1} + \underbrace{e}_{Tx1} \tag{8}$$

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

Solution 11 (CLRM with MATLAB)**.**
See question one solution in MATLABps1sol.pdf.

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:

$$\triangle p_{it} = \phi_{0i} + \phi_{1i} p_{it-1} + \phi_{2i} \triangle p_{it-2} + e_{it}, e_{it} \sim (0, \sigma_i^2). \tag{9}$$

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:

$$\triangle d_{it} = \delta_{0i} + \delta_{1i} e_{it-1} + \delta_{2i} e_{it-2} + e_{it}, e_{it} \sim iidN(0, \sigma_i^2). \tag{10}$$

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:

$$\triangle p_{it} = \phi_{0i} + \phi_{1i} \triangle p_{it-1} + \delta_{1i} e_{it-1} + e_{it}, e_{it} \sim iidN(0, \sigma_i^2). \tag{11}$$

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,

$$y_t = \phi_0 + \phi_1 yt - 1 + \phi_2 y_{t-2} + e_t, e_t \sim (0, \sigma_e^2). \tag{12}$$

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)

Solution 12 (ARMA with MATLAB)**.**
See question two solution in MATLABps1sol.pdf.

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)

Solution 13 (Distributed lag model)**.**
In the general distributed lag model

$$A(L)y_t = B(L)x_t + u_t$$

where $A(L) = 1 + \alpha_1 L + \alpha_2 L^2 + \cdots$ and $B(L) = \beta_0 + \beta_1 L + \beta_2 L^2 + \cdots$, various quantities of interest can be extracted from the polynomials $A(L)$ and $B(L)$. Consider the mean path of y_t

$$\bar{y}_t = E(y_t) = \frac{B(L)}{A(L)} x_t = D(L)x_t = \delta_0 x_t + \delta_1 x_{t-1} + \delta_2 x_{t-2} + \cdots \tag{13}$$

Suppose that x_t is in equilibrium and hence, $x_t = x$. The total multiplier is the effect that a unit change in x has on the equilibrium value of $\bar{y}_t$ and is defined as

$$\text{Total multiplier } = \delta_0 + \delta_1 + \delta_2 + \cdots = D(1) = \frac{B(1)}{A(1)}$$

The mean lag, on the other side, is defined as

$$\text{Mean lag } = \frac{\sum_{j=0}^{\infty} \delta_j j}{\sum_{j=0}^{\infty} \delta_j} = \frac{\partial D(L)}{\partial L}\bigg|_{L=1} \left[\frac{1}{D(1)}\right] = \frac{D'(1)}{D(1)} = \frac{B'(1)}{B(1)} - \frac{A'(1)}{A(1)}$$

Parts **(a)** and **(b)**. In the mean path

$$\bar{y}_t = \frac{3L}{1 - 0.9L + 0.2L^2} x_t \tag{14}$$

we have that $A(L) = 1 - 0.9L + 0.2L^2$ and $B(L) = 3L$, so $A(1) = 1 - 0.9 + 0.2 = 0.3$, $A'(L) = -0.9 + 0.4L$, $A'(1) = -0.5$ and $B(1) = B'(L) = B'(1) = 3$. Thus,

$$\text{Total multiplier } = \frac{B(1)}{A(1)} = \frac{3}{0.3} = 10$$

and

$$\text{Mean lag } = \frac{B'(1)}{B(1)} - \frac{A'(1)}{A(1)} = \frac{3}{3} + \frac{0.5}{0.3} = 2.67$$

Part **(c)**. We are now interested in finding the coefficients δ_j in (13), using the structure given in (14). Note first that $A(L) = 1 - 0.9L + 0.2L^2 = (1 - \mu_1 L)(1 - \mu_2 L)$ where $\mu_1 = 0.4$ and $\mu_2 = 0.5$. We now use partial fractions to expand $B(L)|A(L)^{-}1|$. Write this polynomial as

$$\frac{B(L)}{A(L)} = \frac{3L}{(1 - \mu_1 L)(1 - \mu_2 L)} = \frac{w_1}{1 - \mu_1 L} + \frac{w_2}{1 - \mu_2 L}$$

$$\text{so } 3L = w_1(1 - \mu_2 L) + w_2(1 - \mu_1 L) = (w_1 + w_2) - (w_1\mu_2 + w_2\mu_1)L \longrightarrow \begin{cases} w_1 + w_2 = 0 \\ w_1\mu_1 + w_2\mu_2 = -3 \end{cases}$$

Thus, $w_1 = -w_2 = \frac{3}{\mu_1 - \mu_2} = 30$. With this result, (14) becomes

$$\begin{aligned} \bar{y}_t &= \left(\frac{30}{1 - 0.5L} - \frac{30}{1 - 0.4L} \right) x_t \\ &= \frac{30}{1 - 0.5L} x_t - \frac{30}{1 - 0.4L} x_t \\ &= 30 \sum_{j=0}^{\infty} [(0.5)^j - (0.4)^j] x_{t-j} \end{aligned} \tag{15}$$

Thus, we get that $\delta_j = 30[(0.5)^j - (0.4)^j]$ – compare (15) with (13) – so

$$\begin{aligned} \delta_0 &= 30[(0.5)^0 - (0.4)^0] = 0 \\ \delta_1 &= 30[(0.5)^1 - (0.4)^1] = 30 \cdot 0.1 = 3 \\ \delta_2 &= 30[(0.5)^2 - (0.4)^2] = 30[0.25 - 0.16] = 30 \cdot 0.09 = 2.7 \end{aligned}$$

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?

Solution 14 (Reparameterisation)**.**
Consider the following regression

$$y_t = \delta_0 x_t + \delta_1 x_{t-1} + \delta_2 x_{t-2} + \epsilon_t \tag{16}$$

To reparameterise it, add and subtract $\delta_0 x_{t-1}$ and then add and subtract $(\delta_0 + \delta_1)x_{t-2}$, so

$$\begin{aligned} y_t &= \delta_0 x_t - \delta_0 x_{t-1} + \delta_0 x_{t-1} + \delta_1 x_{t-1} + \delta_2 x_{t-2} + \epsilon_t \\ &= \delta_0 \Delta x_t + (\delta_0 + \delta_1)x_{t-1} + \delta_2 x_{t-2} + \epsilon_t \\ &= \delta_0 \Delta x_t + (\delta_0 + \delta_1)x_{t-1} - (\delta_0 + \delta_1)x_{t-2} + (\delta_0 + \delta_1)x_{t-2} + \delta_2 x_{t-2} + \epsilon_t \\ &= \delta_0 \Delta x_t + (\delta_0 + \delta_1)\Delta x_{t-1} + (\delta_0 + \delta_1 + \delta_2)x_{t-2} + \epsilon_t \end{aligned} \tag{17}$$

Define $\delta_0 * = \delta_0$, $\delta_1 * = \delta_0 + \delta_1$ and $\delta_2 * = \delta_0 + \delta_1 + \delta_2$ so (17) becomes

$$y_t = \delta_0 * \Delta x_t + \delta_1 * \Delta x_{t-1} + \delta_2 * x_{t-2} + \epsilon_t \tag{18}$$

It is clear that the new parameters $\delta_i *$ capture the accumulated effect of a uniut change in x_t on y_t over time (i.e., the interim multipliers). Hence, the parameter associated with the furthest lag, $\delta_2 *$ is the long-run response of y_t to the unit shock, or the change between different steady states of y due to the shock in the equilibrium value of x.

Equation (18) has two advantages over (16): if the quantities of interest are the accumulated impulse response values, estimating (18) not only redner those figures directly but also their standard errors. Besides, if (as usual) the x_t's are strongly serially correlated the regression in (16) may suffer from a multicollinearity problem, whereas multicollinearity will not be an issue in (18).

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?

Solution 15 (Forecasting an MA(1) process)**.**
The best j-step ahead linear predictor of a time series is its mean (expectation) conditional on the information available at time T, $\mathbf{Y}_T$, $y_{T+j|T} = E(y_{T+j}|\mathbf{Y}_T) = E_T(y_{T+j})$.

Consider the MA(1) process $y_t = \alpha + \epsilon_t + \theta\epsilon_{t-1}$, where the last realisation of ϵ_t is ϵ_T implying that $E(\epsilon_s|\mathbf{Y}_T) = E(\epsilon_s) = 0$ for any $s > T$ and $E(\epsilon_s|\mathbf{Y}_T) = \epsilon_s$ for $s \leq T$. Then, since

$$y_{T+j} = \alpha + \epsilon_{T+j} + \theta\epsilon_{T+j-1}$$

we have that $y_{T+j|T} = \alpha + E_T(\epsilon_{T+j}) + \theta E_T(\epsilon_{T+j-1})$, so that

$$y_{T+1|T} = \alpha + \theta\epsilon_T \quad \text{and} \quad y_{T+j|T} = \alpha \quad \text{for } j > 1$$

The mean squared error of the predictor is defined as $MSE(y_{T+j|T}) = E_T[(y_{T+j|T} - y_{T+j})^2]$, so for the MA(1) process this is equal to

$$\begin{aligned} MSE(y_{T+1|T}) &= E_T(\epsilon_{T+1}^2) = \sigma^2 \\ MSE(y_{T+j|T}) &= E_T(\epsilon_{T+j}^2) + 2\theta E_T(\epsilon_{T+j}\epsilon_{T+j-1}) + \theta^2 E_T(\epsilon_{T+j-1}^2) = \sigma^2(1+\theta^2) \text{ for } j > 1 \end{aligned}$$

Finally, using $\alpha = 5, \theta = 0.5, \epsilon_T = 1.2$ and $\sigma^2 = 1$ we get $y_{T+1|T} = 5.6$, $MSE(y_{T+1|T}) = 1$ and $y_{T+2|T} = 5$, $MSE(y_{T+2|T}) = 1.25$.

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

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

Solution 16 (Forecasting an ARMA(2,2) process)**.**
As in the previous question, the best linear predictor is the conditional mean. Since

$$y_{T+1} = 0.6y_T + 0.2y_{T-1} + \epsilon_{T+1} + 0.6\epsilon_T$$

and $\{y_1, y_2, \ldots, y_{T-1}, y_T\}$ and ϵ_T are known at time T, we have (for $y_T = 2.0$, $y_{T-1} = 1.0$ and $\epsilon_T = 0.5$) that

$$y_{T+1|T} = E_T(y_{T+1}) = 0.6y_T + 0.2y_{T-1} + 0 + 0.6\epsilon_T = 1.7$$

Similarly

$$y_{T+2|T} E_T(y_{T+2}) = 0.6y_{T+1|T} + 0.2y_T = 1.42$$
$$y_{T+3|T} E_T(y_{T+3}) = 0.6y_{T+2|T} + 0.2y_{T+1|T} = 1.192$$

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.

Solution 17 (Minimum MSE Forecasts)**.**
Given a loss function $L(e) = e^2$, where $e = Y - f$ we want to show that $E(L(e))$ is the MSE.

$$E(L(e)) = E(e^2) = E(e^2) = E[(Y - f)^2]$$

$$\begin{aligned} E[(Y - f)^2] &= E\{[(Y - E(Y)) + (E(Y) - f)]^2\} \\ &= E[(Y - E(Y))^2] + E[(E(Y) - f)^2] + 2E[(Y - E(Y))(E(Y) - f)] \end{aligned}$$

Note that $E(Y) - f$ is a constant so $E[(E(Y) - f)^2] = (E(Y) - f)^2$, which is the square of the bias. As $E(Y) - f$ is a constant

$$E[(Y - E(Y))(E(Y) - f)] = [E(Y) - f]\underbrace{E[Y - E(Y)]}_{0} = 0$$

$$\therefore E[(Y - f)^2] = Var(Y) + Bias^2 = MSE$$

So, we have shown that a quadratic loss function implies that associated risk will be the mean square error.

Observe that the minimum MSE forecast will be given by $f = E(Y)$ since this makes the bias equal to zero and thereby minimises the MSE.

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)

Solution 18 (Forecasting: Estimation, Assessment & Using Many Predictors)**.**

1. See section 5.2.3 in the notes.

2. See section 5.2.4 in the notes.

3. See section 5.3 in the notes.