The University Of Sheffield.

PLEASE LEAVE THIS EXAM PAPER ON YOUR DESK.
DO NOT REMOVE IT FROM THE HALL.

Data Provided:
Neaves Tables
Graph Paper

SCHOOL OF MATHEMATICS AND STATISTICS MAS6061

Session 2012-2013 **3 Hours**

Epidemiology and Time Series

RESTRICTED OPEN BOOK EXAMINATION.
Candidates may bring to the examination lecture notes and associated lecture material (but no textbooks) plus a calculator that conforms to University regulations.
*All answers will be marked but credit will be given for only the best **FIVE** answers.*
All questions carry equal marks. Total marks 100.

Registration number from U-Card (9 digits) – to be completed by student

(This page is left blank)

1. Campbell et al (2012) described a tool, the Summary Hospital Mortality Index (SHMI) for predicting death in hospital. It is based on age, sex, diagnosis and method of admission (emergency or not). Overall 2% of patients die in hospital. Suppose 12 patients who died and 12 who did not die were assessed using the SHMI with the following results:

SHMI (x100):

Patients who died: 0.5, 1.3, 1.4, 1.6, 2.5, 3.0, 3.3, 3.5, 4.0, 4.1, 5.5, 6.2

Patients who did not die: 0.1, 0.1, 0.2, 0.4, 0.6, 1.2, 1.3 ,1.5, 1.7, 1.8, 2.0, 2.6

(i) Estimate the odds ratio of dying with a SHMI >1.1 compared to a SMHI less than or equal to 1.1. Can we use this odds ratio as an estimate of the relative risk of dying given a SHMI over 1.1?

(2 marks)

(ii) Find the sensitivity of the predictor for a specificity of 75%.

(2 marks)

(iii) Plot the ROC using the three quartiles of the sample distribution of SHMI as cut-off points.

(4 marks)

(iv) Find the area under the ROC curve found in (iii) and interpret it.

(6 marks)

(v) What statistics would be useful to an individual patient to decide on their risk of dying and can you derive them from this table. If not, why not?

(2 marks)

(vi) What other factors have to be in place for the SHMI to be a useful tool for patients to decide whether a hospital carries a higher risk of death for them?

(4 marks)

2.

(i) Prevalence and incidence are both measures of disease etiology. Explain the difference between these two measures.

(4 marks)

(ii) In a recent study of asthma trends in schoolchildren in Scotland the prevalence of asthma in 1999 was 24.3% (in a sample of n=3540 children) and in 2009 it was 22.1% (n=1103). Display these data in a 2 by 2 table and calculate a suitable measure for comparing asthma prevalence between the two years, together with its 95% confidence interval. Do the results suggest that the prevalence has changed between the two years?

(7 marks)

(iii) These data were further broken down by gender. Using a suitable procedure calculate the odds ratio of asthma for 2009 compared to 1999, after having adjusted for the effect of gender. Please comment on the result.

(8 marks)

Table 2(a): Asthma prevalance: Boys

	1999	2009
Asthma	456	139
No asthma	1,296	396

Table 2(a): Asthma prevalance: Girls

	1999	2009
Asthma	404	104
No asthma	1,384	463

(iv) Is there evidence of a significant difference in the gender adjusted odds of asthma for 2009 compared to 1999?

(1 mark)

3. A study was conducted in an attempt to replicate a previously reported association between risk of Parkinson's disease (PD) and genetic variants in the *VMAT2* gene. The previous report found evidence of a dominant effect associated with the rare alleles at two SNPs. The same two SNPs have been genotyped in an Italian series of 704 PD patients and 678 unrelated healthy controls. The genotyping results are shown in the table below.

SNP		Common homozygote	Heterozygote	Rare homozygote
rs363371 (A/G)	Cases Controls	GG: 464 GG: 396	GA: 208 GA: 245	AA: 31 AA: 36
rs363324 (A/G)	Cases Controls	AA: 364 AA: 304	AG: 259 AG: 291	GG: 70 GG: 74

(i) Calculate the allele frequency for allele A in controls for the SNP rs363371.

(1 mark)

(ii) Calculate the expected genotype frequencies of the controls for SNP rs363371 under Hardy Weinberg Equilibrium (HWE). Is there evidence against HWE in the controls for this SNP?

(3 mark)

(iii) The appropriate statistical test for deviation from HWE returns a p-value of 0.73 for the SNP rs363324. Genetic association studies frequently set a low type 1 error threshold (such as 1×10^{-4}) when testing for HWE. Why is this and what is the conclusion if significant evidence is found?

(2 marks)

(iv) Assess the evidence for a fully dominant effect of the rare alleles for both SNPs on risk of PD. In each case use the most appropriate test with 1 degree of freedom, and report the summary comparative statistic.

(6 marks)

(v) The two SNPs are located within the same gene and linkage disequilibrium (LD) has been measured between them as follows: r^2=0.61, D'=1.00. Discuss the results you have generated in (iv) in the light of this knowledge of the pairwise LD.

(3 marks)

(vi) Parkinson's disease risk is known to be associated with smoking and gender and PD risk is higher in older subjects. The study had measured these three variables in the cases and controls. What statistical method would be appropriate to allow for these known risk factors and also test for a dominant association between these risk SNPs and PD.

(1 mark)

(vii) The report was also interested in testing if the genetic variants in VMAT2 were associated with severity of PD. In this case more severe disease was judged as having an earlier age at onset. Outline a method to test this hypothesis naming an appropriate statistical test.

(4 marks)

4 The plot above shows quarterly total sales (in thousands) of one-day-old turkey chicks from hatcheries in Eire (source Pole, A., West, M., and Harrison, P.J., 1994, Applied Bayesian Forecasting and Time Series Analysis, Chapman-Hall).

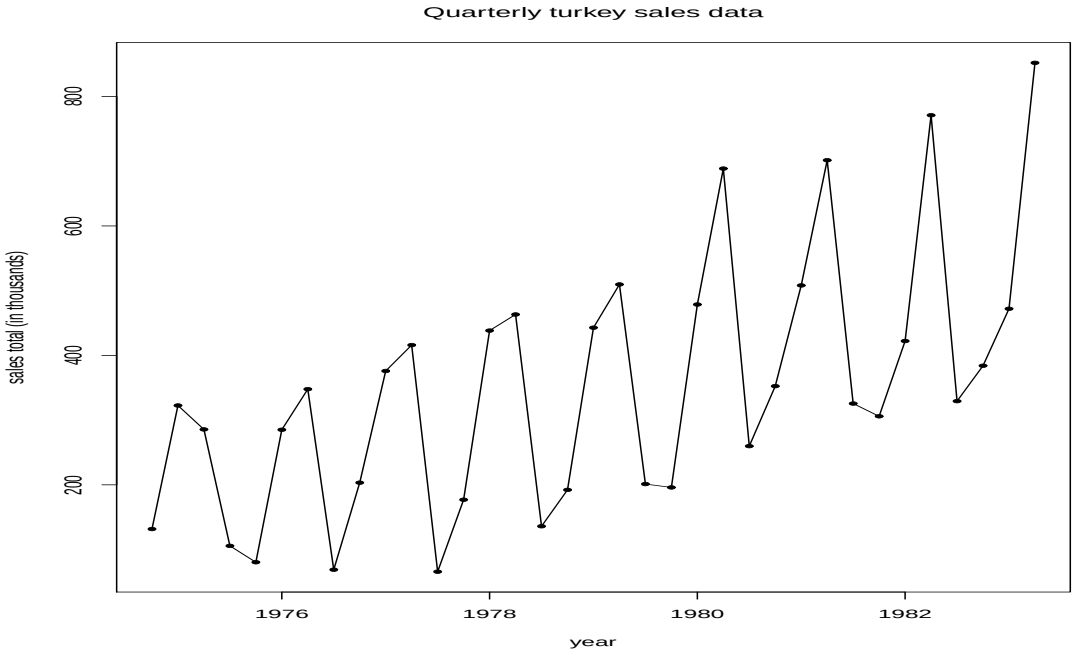

Figure 1: Quarterly sales turkey chicks data

(i) Briefly describe the features of the data. ***(2 marks)***

(ii) Suggest a transformation of the time series data y_t, likely to result in a stationary time series x_t and write down x_t as a function of y_t using appropriate differencing notation. ***(2 marks)***

(iii) For a time series x_t (of length 31) derived from y_t by a suitable transformation, the sample ACF and the sample PACF are tabulated below:

Lag	1	2	3	4	5	6	7	8
ACF	r_1	r_2	-0.468	0.700	-0.397	0.134	0.051	0.002

and

Lag	1	2	3	4	5	6	7	8
PACF	-0.650	-0.488	-0.832	-0.565	-0.300	0.124	0.090	0.032

(a) Find the values of r_1 and r_2. ***(4 marks)***

(b) Test whether x_t is a white noise. ***(2 marks)***

(c) Test whether x_t is consistent with autoregressive models. ***(3 marks)***

(d) Test whether x_t is consistent with moving average models. ***(5 marks)***

(e) Based on your analysis above, suggest a time series model for x_t that is likely to perform well when fitted to the data. ***(2 marks)***

5 Consider the time series model

$$y_t = 19 - \frac{1}{3}y_{t-1} - \frac{1}{4}y_{t-2} + \epsilon_t - \frac{1}{2}\epsilon_{t-1},$$

where ϵ_t is white noise with variance 8.

(i) Write down this model using the Backward shift operator B. ***(2 marks)***

(ii) Show that this model is causal and invertible. ***(5 marks)***

(iii) Find the mean of y_t. ***(3 marks)***

(iv) Find the variance of y_t. ***(10 marks)***

6 The plot below relates y_t the quarterly change in a company's sales to x_t the quarterly change of a market sales indicator variable.

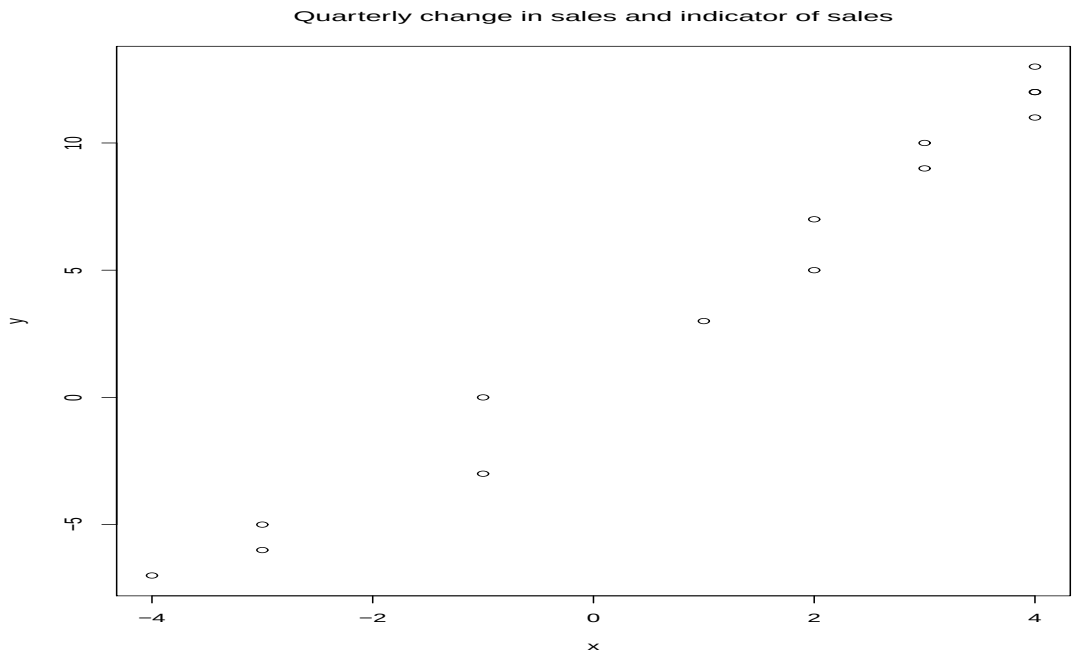

Figure 2: Relation of change of sales y_t and indicator of sales x_t

As a first model it is suggested to regress y_t on x_t, or

$$y_t = x_t\beta + \epsilon_t,$$

where β is a regression coefficient and ϵ_t is a white noise with variance 1.

However, the statistician of the company argues this model is not appropriate to model the data set. To back her argument she has provided the autocorrelation functions of the time series x_t and y_t, given in the plot overleaf (Figure 3).

(i) Explain why the statistician believes the model above is inappropriate.

(1 mark)

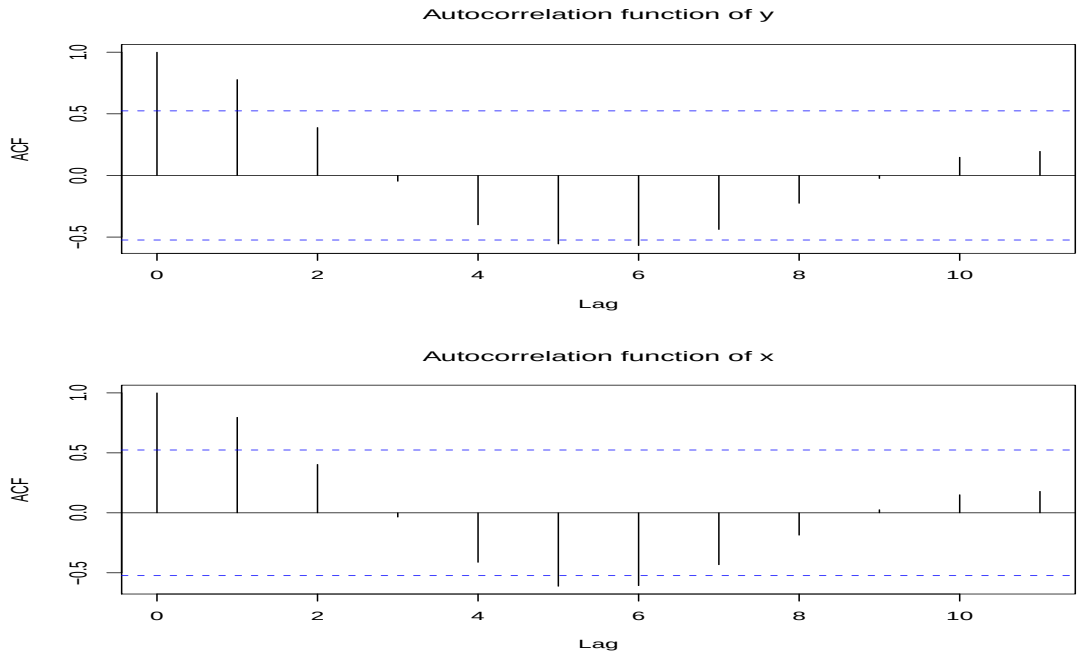

Figure 3: ACF of the time series x_t and y_t

6 (continued)

(ii) The statistician suggested a second model as alternative, given by

$$y_t = x_t\beta_t + \epsilon_t \quad \text{and} \quad \beta_t = \beta_{t-1} + \zeta_t, \tag{1}$$

where ϵ_t is as before and ζ_t is a white noise with variance 10.

(a) Give the name of model (1). ***(1 mark)***

(b) For model (1) show that $P_{t|t}$ the posterior variance of β_t satisfies

$$\frac{1}{P_{t|t}} = \frac{1}{P_{t-1|t-1} + 10} + x_t^2.$$

(7 marks)

(c) If $x_1 = 4$, $x_2 = 4$, $y_1 = 12$, $y_2 = 11$ and the prior of β_0 is $\beta_0 \sim N(2, 0.81)$, then use the result in (b) to calculate the posterior means $\hat{\beta}_{1|1}$, $\hat{\beta}_{2|2}$ and the posterior variances $P_{1|1}$, $P_{2|2}$. ***(5 marks)***

(d) If $x_3 = 3$, use (c) to obtain the one-step forecast mean of $y_3 = 9$ and the associated residual. Comment on the quality of this forecast. ***(2 marks)***

6 (continued)

(e) If instead of $\beta_0 \sim N(2, 0.81)$ the prior distribution of β_0 is set to either of the following:

(α) $\beta_0 \sim N(10, 0.81)$ or

(β) $\beta_0 \sim N(2, 100)$,

comment on whether you expect an improvement on forecasting and the general model performance for all y_t. ***(3 marks)***

(f) Suggest how the model performance can be improved. ***(1 mark)***

End of Question Paper