Name: _____UMID #: _____

Midterm Exam

There are **3 sections (A, B and C) containing a total of 10 questions worth 23 points.**
Points will be awarded for clearly explained and accurate answers. You may use the
course textbook, the course notes and a calculator.

Section	Points	Score
A	7	
B	8	
C	8	
Total	23	

We investigate the monthly number mumps cases reported in New York City, from January 1928 to June 1972. During this period, before the introduction of a vaccine, mumps was a common childhood disease to which almost all children were exposed. Because mumps displays a characteristic rash it is fairly easily diagnosed. Mumps is a reportable disease, meaning that doctors have a legal obligation to report any cases they encounter. This dataset therefore gives an opportunity to study disease transmission and maybe learn lessons relevant to diseases of current concern such as bird flu, SARS or HIV/AIDS. The data, which we shall denote by $\{x_t, t = 1, 2, \ldots\}$, are graphed in Fig. 1.
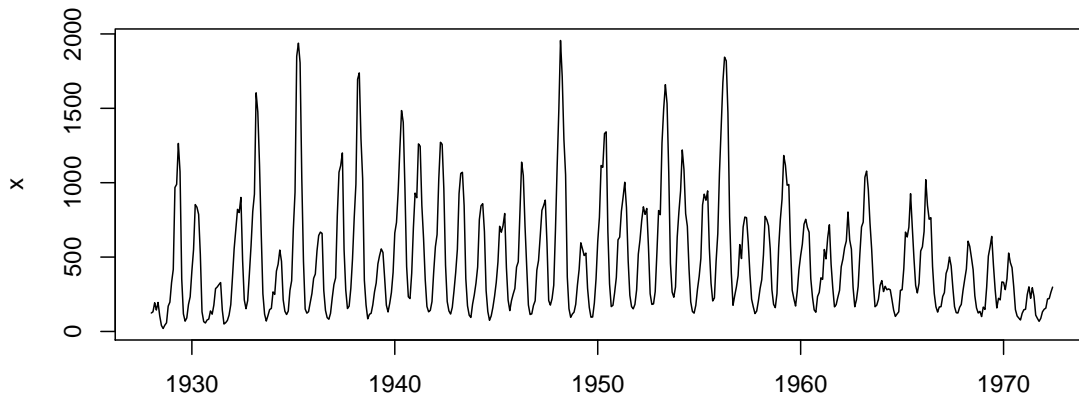

Figure 1: Monthly mumps reports, x_t, in New York City from January 1928 to June 1972.

Section A. Spectral analysis. We seek to interpret the estimated spectum in Fig. 2 and in particular the features labeled (1) through (5).
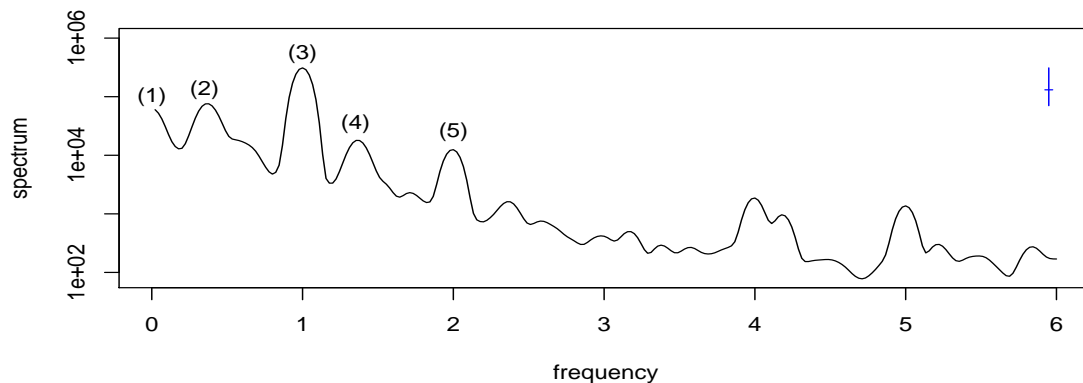

Figure 2: An estimated spectral density for x_t, calculated via `spectrum(x,spans=c(3,5,7))`.

A1. [2 points] What are the units of frequency in Fig. 2? Explain how you reach your answer.

2

A2. [3 points] One might expect mumps to have annual seasonality. One might also expect mumps to have long term cycles as the population of suscpetible children (those without immunity) replenishes after previous outbreaks. Discuss the interpretation of the 5 spectral peaks labeled (1) through (5) in Fig. 2. You do not have to discuss here whether these peaks are statistically significant, which is question A3 below.

A3. [2 points] Comment on the statistical significance of these five peaks. You are not expected to present formal tests, but you should say what your opinion is and why.

Section B. ARIMA analysis. We try fitting an $ARIMA(3,0,0) \times (0,1,1)_{12}$ model. Call this model M1. The output from `M1=arima(x,order=c(3,0,0),seasonal=c(0,1,1))` is

```
         ar1      ar2      ar3     sma1
      1.2032  -0.3025  -0.0632  -0.8841
s.e.  0.0439   0.0674   0.0442   0.0231

sigma^2 estimated as 12881:  log likelihood = -3220.76,  aic = 6451.51
```

Another possibility is to model $\log(x_t)$, again using $ARIMA(3,0,0) \times (0,1,1)_{12}$. Call this model M2. The output from `M2=arima(log(x),order=c(3,0,0),seasonal=c(0,1,1))` is

```
         ar1     ar2      ar3     sma1
      0.9197  0.1577  -0.1710  -0.8080
s.e.  0.0434  0.0592   0.0438   0.0285

sigma^2 estimated as 0.03632:  log likelihood = 117.48,  aic = -224.96
```

B1. [2 points] Can the above analysis determine whether a log transformation is appropriate? Explain.

A table comparing AIC values for various $ARIMA(i,0,j) \times (0,1,1)_{12}$ models for $\log(x_t)$ is given below:

AR \ MA	0	1	2	3	4
0	NA	312.7628	92.7453	-42.91403	-131.8598
1	-213.9453	-211.9458	-224.4315	-227.35447	-225.5215
2	-211.9459	-212.5350	-236.8260	-223.70594	-224.7305
3	-224.9618	-237.9834	-236.1537	-234.41349	-232.4061
4	-229.8224	-221.1222	-236.4941	-235.21621	-239.7320

B2. [2 points] The software gave no error messages while computing this table. Is there any reason to suspect that the numeric maximization of the likelihood is less than adequate?

B3. [4 points] Discuss briefly what you learn from the AIC table shown, in terms of developing a suitable model for these data. Explain briefly why AIC may not be the only criterion considered when selecting a model, and list some other analyses that you would carry out to determine and defent a choice of model.

Section C. Diagnostic analysis. Fig. 3 contains six diagnostic plots, three for each of models M1 and M2.

C1. [2 points] Explain carefully the meaning of the dashed line in sample ACF plots produced by R, for example in Fig. 3(a1). [Here, you are asked to explain the statistical method; later parts will ask you to interpret the results in the context of the data and models under investigation.]
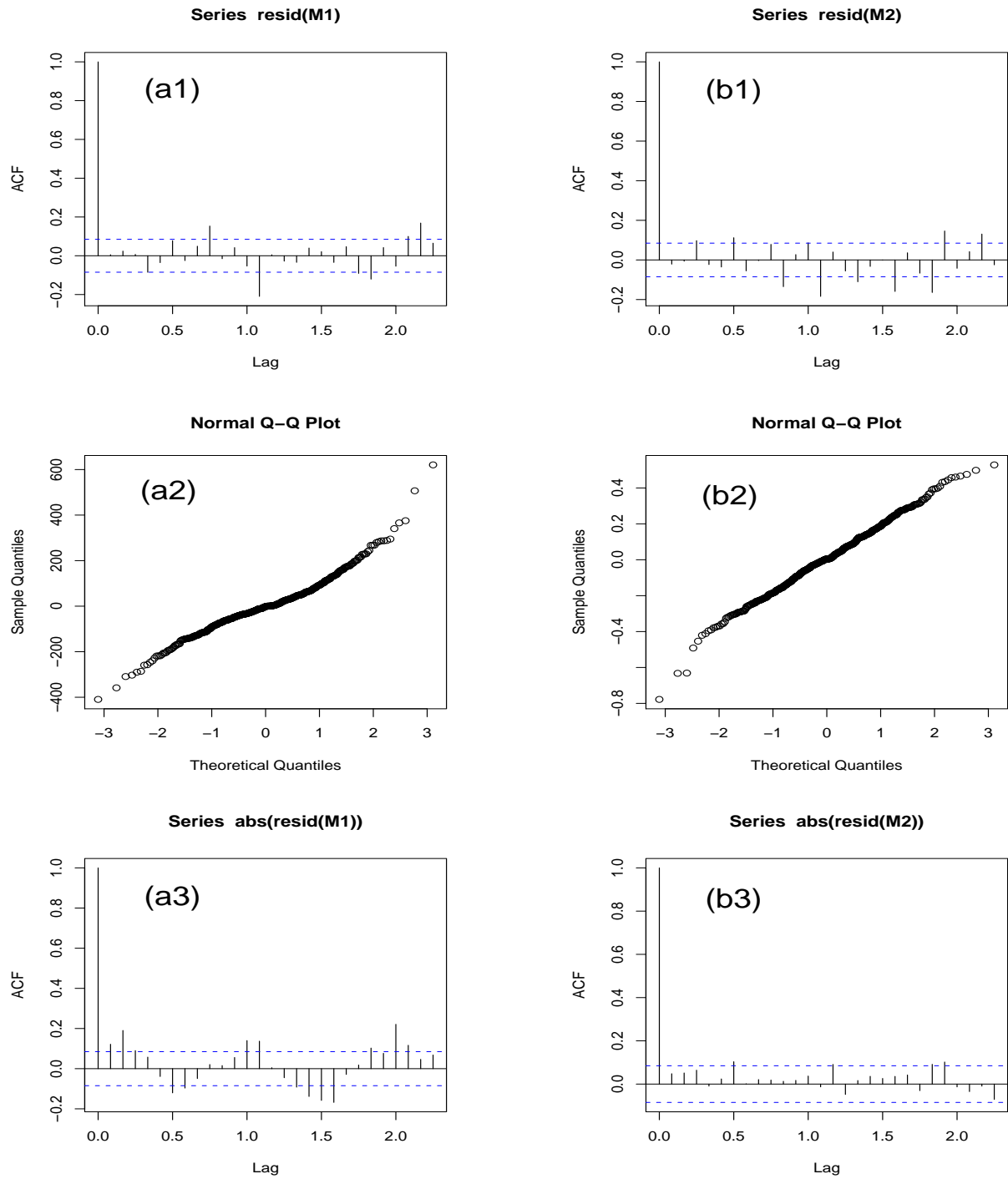
Figure 3: Some diagnostic plots. (a1) and (a3) show the sample ACF for the residuals and absolute values of the residuals respectively for model M1. (a2) is a normal quantile plot of the residuals for M1. (b1,b2,b3) are the equivalent diagnostic plots for M2.

C2. [2 points] Compare (a1) and (b1) in Fig. 3. What does this tell you about models M1 and M2?

C3. [2 points] Compare (a2) and (b2) in Fig. 3. What does this tell you about models M1 and M2?

C4. [2 points] Compare (a3) and (b3) in Fig. 3. What does this tell you about models M1 and M2? In particular, what do you learn about the appropriateness of an assumption that the white noise process driving the ARIMA model is independent and identically distributed?