Midterm 1, STATS 531/631 W26

In class on 2/16

Name: **UMID:**

This document produces different random tests each time the source code generating it is run. The actual midterm will be a realization generated by this random process, or something similar.

This version lists all the questions currently in the test generator. The actual test will have one question sampled from each of the 7 question categories.

Instructions. The test is closed book, and you are not allowed access to any notes. Any electronic devices in your possession must be turned off and remain in a bag on the floor.

For each question, circle one letter answer and provide some supporting reasoning.

Q1. Stationarity and unit roots.

Q1-01.

Suppose that a dataset $y^*_{1:N}$ is well described by the statistical model

$$Y_n = a + bn + \epsilon_n,$$

where ϵ_n is a Gaussian ARMA process and $b \neq 0$. Which of the following is the best approach to time series modeling of $y^*_{1:N}$?

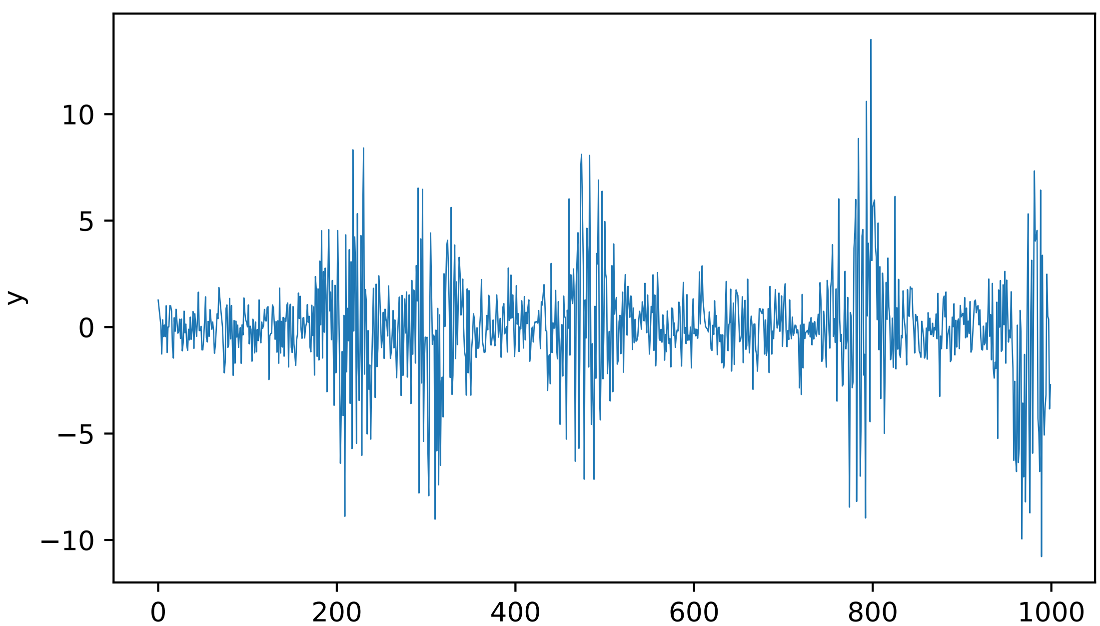
A. The data are best modeled as non-stationary, so we should take differences. The differenced data are well described by a stationary ARMA model.

B. The data are best modeled as non-stationary, and we should use a trend plus ARMA noise model.

C. The data are best modeled as non-stationary. It does not matter if we difference or model as trend plus ARMA noise since these are both linear time series models which become equivalent when we estimate their parameters from the data.

D. We should be cautious about doing any of A, B or C because the data may have nonstationary sample variance in which case it may require a transformation before it is appropriate to fit any ARMA model.

Q1-02.



Consider the time series plotted above. Which of the below is the most accurate statement about stationarity?

A. The plot shows that the data are clearly non-stationary. We could make a formal hypothesis test to confirm that, but it would not be insightful. To describe the data using a statistical model, we will need to develop a model with non-constant variance.

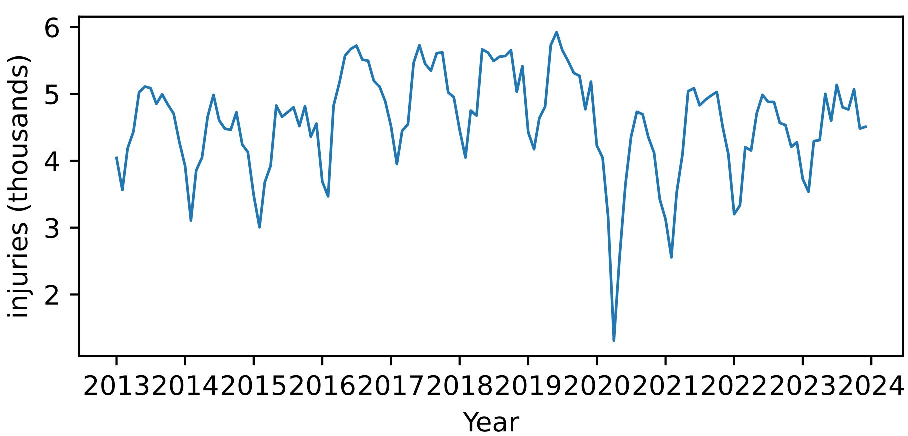
B. The sample variance is evidently different in different time intervals. However, we should not conclude that the underlying data generating mechanism is non-stationary before making a formal statistical test of equality of variances between the time regions that have lower sample variance and the regions that have higher sample variance. Visual impressions without a formal hypothesis test can be deceptive.

C. A model with randomly changing variance looks appropriate for these data. Since the variance for such a model is time-varying, the model must be non-stationary.

D. A model with randomly changing variance looks appropriate for these data. Despite the variance for such a model being time-varying, the model is stationary.

E. The sample variance is evidently different in different time intervals. An appropriate next step to investigate stationary would be to plot the sample autocorrelation function for different intervals to see if the dependence between time points is also time-varying.

Q1-03.



Above are monthly injuries from motor vehicle collisions in New York City. An augmented Dickey-Fuller test, `adfuller(injuries)`, gives a p-value of 0.014528. Which is the best way to proceed:

A: The time plot indicates a non-constant mean function describing a major dip due to the COVID-19 pandemic and an increasing trend at other times. The ADF test does not support or refute that model.

B: The ADF test suggests the series is stationary, supporting a decision to fit a SARMA model.

C: The ADF test suggests the series is non-stationary; it should be differenced before fitting a SARMA.

D: The ADF test indicates that the series is non-stationary, supporting the use of a non-constant mean function to describe a major dip due to the COVID-19 pandemic and an increasing trend at other times.

Q2. Calculations for ARMA models

Q2-01.

Let $Y_n = \phi Y_{n-1} + \epsilon_n$ for $n = 1, 2, ...$ with $\epsilon_n \sim \text{iid}N[0, \sigma^2]$ and $Y_0 = 0$. The covariance of Y_n with Y_{n+k} for $k \geq 0$ is

A. $\sigma^2 \phi^k / (1 - \phi^2)$

B. $\sigma^2 \phi^{2k} / (1 - \phi^2)$

C. $\sigma^2 \phi^k / (1 - \phi)$

D. $\sigma^2 \phi^{2k} / (1 - \phi)$

E. None of the above.

Q2-02.

Let Y_n be an ARMA model solving the difference equation

$$Y_n = (1/4)Y_{n-2} + \epsilon_n + (1/2)\epsilon_{n-1}.$$

This is equivalent to which of the following:

A. $Y_n = (1/2)Y_{n-1} + \epsilon_n$

B. $Y_n = -(1/2)Y_{n-1} + \epsilon_n$

C. $Y_n = (1/2)Y_{n-2} - (1/16)Y_{n-4} + \epsilon_n + \epsilon_{n-1} + (1/4)\epsilon_{n-2}$

D. $Y_n = -(1/2)Y_{n-2} - (1/16)Y_{n-4} + \epsilon_n + \epsilon_{n-1} + (1/4)\epsilon_{n-2}$

E. None of the above

Q2-03.

Is it possible for an $AR(2)$ model to have a finite moving average representation, so that it is equivalent to some $MA(q)$ model for $q < \infty$?

A. No. Any moving average representation of any $AR(2)$ model is $MA(\infty)$

B. Yes. Although it is not true for any $AR(2)$ process, it is possible to find particular choices of the autoregressive coefficients, p_1 and p_2, that lead to a finite $MA(q)$ representation.

C. It is not possible for any real-valued p_1 and p_2, but it is possible if you permit p_1 and p_2 to be complex-valued.

Q2-04.

Different criteria for selecting a time series model include (i) Akaike's Information Criterion; (ii) leave-one-out cross-validation; (iii) out-of-sample k-step-ahead prediction error; (iv) holding out the most recent 20% of the data for testing, while fitting to the first 80%. Suppose our goal is to make predictions for a collection of forecasting windows, and that the model fits the data fairly well. Which of the following are correct:

A. AIC is based on one-step prediction, so for longer-term forecasts it is less reliable than fitting by k-step prediction error for $k > 1$.

B. Leave-one-out cross-validation is not designed for dependent data.

C. Ideally, we should use a different model for each time in the forecasting window, fitting using k-step prediction error when forecasting k steps ahead.

D. The model that best predicts the most recent data when fitted to earlier data (i.e, (iv)) is more reliable for subsequent forecasting than methods which evaluate based on the whole time series history (i, ii, iii).

E. More than 1 of (A, B, C, D)

Note: You can suppose that the time series model is ARMA, but that is unimportant. All we need is that the model has a likelihood that can be computed, and a conditional expectation that can be evaluated to give a prediction rule.

Q3. Likelihood-based inference for ARMA models

Q3-01.

The following table of AIC values results from fitting ARMA(p,q) models to a time series $y_{1:415}$ where y_n is the time, in miliseconds, between the nth and $(n+1)$th firing event for a monkey neuron. The experimental details are irrelevant here. You are asked to check how many adjacent pairs of AIC values in this table are inconsistent, such that they could mathematically arise only from a numerical error? Adjacent pairs of models are those directly above or below or left or right of each other in the table.

	MA0	MA1	MA2	MA3
AR0	3966.0	3961.5	3962.7	3964.7
AR1	3961.1	3962.6	3964.6	3966.6
AR2	3962.7	3960.5	3959.8	3961.7
AR3	3964.6	3965.5	3962.6	3968.4

A: 0, so the table is mathematically plausible.

B: 1

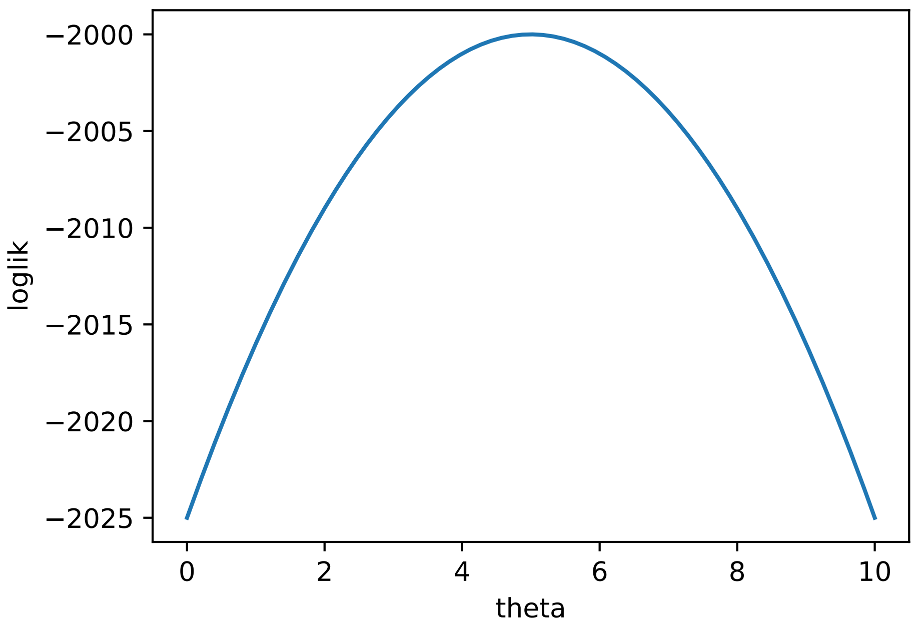
C: 2

D: 3

E: 4 or more

Q3-02.

The Python function `statsmodels.tsa.arima.model.ARIMA.fit` and the R function `arima()` provide standard errors calculated by observed Fisher information. This question tests your understanding of what that means. Suppose a parametric model has a single parameter, θ, and the log-likelihood function when fitting this model to dataset is as follows:



What is the observed Fisher information (I_{obs}) for θ?

Hint 1. The observed Fisher information is accumulated over the whole dataset, not calculated per observation, so we don't have to know the number of observations, N.

Hint 2. Observations in time series models are usually not independent, so the log-likelihood is not the sum of the log-likelihoods for each observation. Its calculation will involve consideration of the dependence, and usually the job of calculating the log-likelihood is left to a computer.

Hint 3. The usual variance estimate for the maximum likelihood estimate, $\hat{\theta}$, is $\text{Var}(\hat{\theta}) \approx 1/I_{obs}$.

A: $I_{obs} = 2$
B: $I_{obs} = 1$
C: $I_{obs} = 1/2$
D: $I_{obs} = 1/4$
E: None of the above

Q3-03.

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  165
Model:                 ARIMA(2, 0, 1)   Log Likelihood                  21.419
Date:                Sun, 08 Feb 2026   AIC                            -32.838
Time:                        22:10:20   BIC                            -17.308
```

```
Sample:                                 0   HQIC                           -26.534
                                    - 165
Covariance Type:                      opg
===============================================================================
```

```
                               SARIMAX Results
===============================================================================
Dep. Variable:                          y   No. Observations:                  165
Model:                     ARIMA(2, 0, 2)   Log Likelihood                  22.634
Date:                    Sun, 08 Feb 2026   AIC                            -33.268
Time:                            22:10:20   BIC                            -14.632
Sample:                                 0   HQIC                           -25.703
                                    - 165
Covariance Type:                      opg
===============================================================================
```

The Python output above uses `ARIMA` from `statsmodels` to fit ARMA(2,1) and ARMA(2,2) models to the January level (in meters above sea level) of Lake Huron from 1860 to 2024. Residual diagnostics (not shown) show no major violation of model assumptions. We aim to choose one of these as a null hypothesis of no trend for later comparison with models including a trend.

Which is the best conclusion from the available evidence:

A: The ARMA(2,2) model has a lower AIC so it should be preferred.

B: We cannot reject the null hypothesis of ARMA(2,1) since the ARMA(2,2) model has a likelihood less than 1.92 log units higher than ARMA(2,1). Since there is not sufficient evidence to the contrary, it is better to select the simpler ARMA(2,1) model.

C: Since the comparison of AIC values and the likelihood ratio test come to different conclusions in this case, it is more-or-less equally reasonable to use either model.

D: When the results are borderline, numerical errors in the `stats::arima` optimization may become relevant. We should check using optimization searches from multiple starting points in parameter space, for example, using `arima2::arima`.

Q3-04.

Suppose model M_0 is nested within a larger model M_1 which has one additional parameter. Suppose that the AIC for M_1 is 0.5 units lower than the AIC for M_0. Which of the following is a correct expression for the p-value of a likelihood ratio test for M_1 against the null hypothesis M_0, supposing that a Wilks approximation is accurate? Here, χ^2_1 is a chi-square random variable on 1 degree of freedom.

A: $P(\chi_1^2 > 0.5)$

B: $P(\chi_1^2 > 1)$

C: $P(\chi_1^2 > 1.5)$

D: $P(\chi_1^2 > 2)$

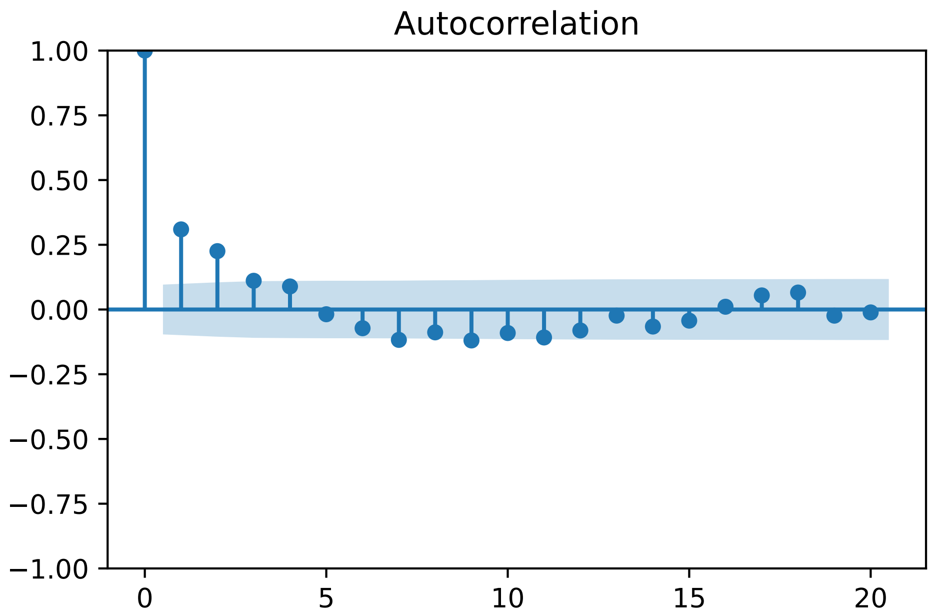
E: $P(\chi_1^2 > 2.5)$

F: $P(\chi_1^2 > 3)$

Q4. Interpreting diagnostics

Q4-01.

We consider data $y_{1:415}$ where y_n is the time, in miliseconds, between the nth and $(n+1)$th firing event for a monkey neuron. Let $z_n = \log(y_n)$, with log being the natural logarithm. The sample autocorrelation function of $z_{1:415}$ is shown below.



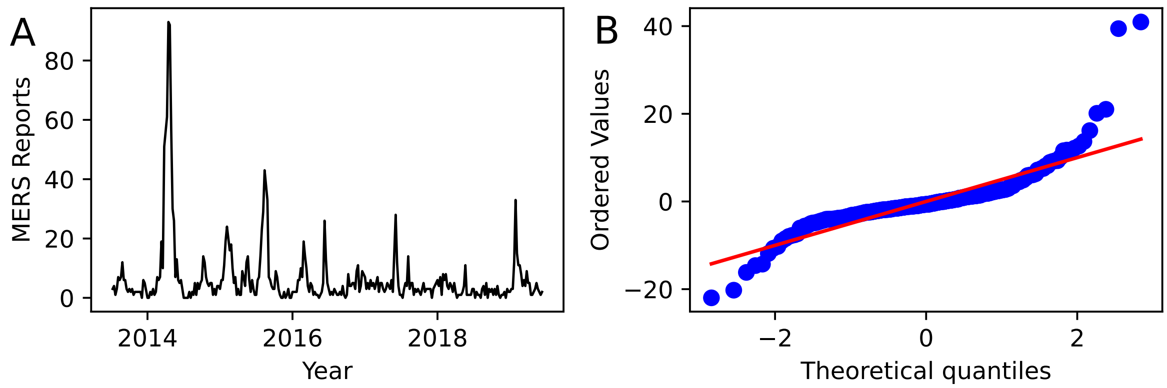
We are interested about whether it is appropriate to model the time series as a stationary causal ARMA process. Which of the following is the best interpretation of the evidence from these plots:

A. There is clear evidence of a violation of stationarity. We should consider fitting a time series model, such as ARMA, and see if the residuals become stationary.

B. This plot suggests there would be no benefit from detrending or differencing the time series before fitting a stationary ARMA model. It does not rule out a sample covariance that varies with time, which is incompatible with ARMA.

C. This plot is enough evidence to demonstrate that a stationary model is reasonable. We should proceed to check for normality, and if the data are also not far from normally distributed then it is reasonable to fit an ARMA model by Gaussian maximum likelihood.

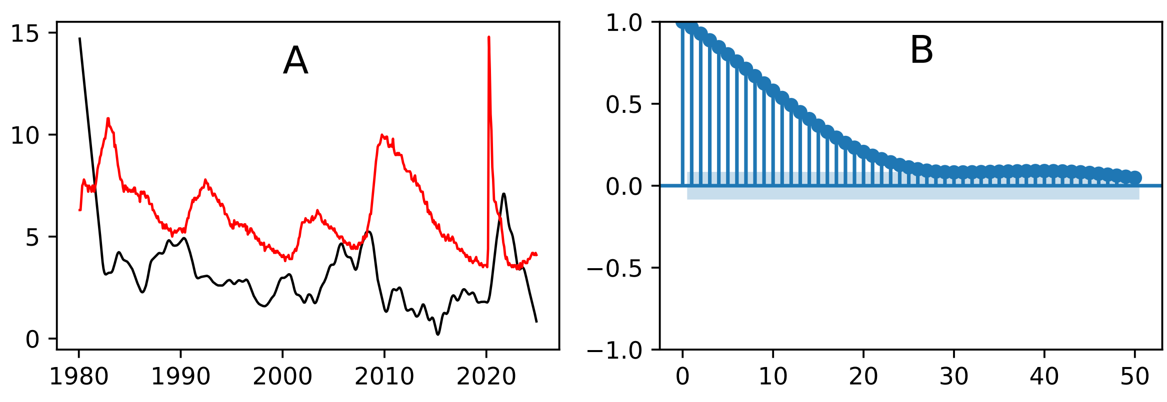
Q4-02.



(A) Weekly cases of Middle East Respiratory Syndome (MERS) in Saudi Arabia. (B) a normal quantile plot of the residuals from fitting an ARMA(2,2) model to these data using `arima()`. What is the best interpretation of (B)?

A: We should consider fitting a long-tailed error distribution, such as the t distribution.
B: The model is missing seasonality, which could be critical in this situation.
C: For using ARMA methods, these data should be log-transformed to make a linear Gaussian approximation more appropriate.
D: The normal quantile plot shows a long-tailed distribution, but this is not a major problem. We have over 300 data points, so the central limit theorem should hold for parameter estimates.
E: The normal quantile plot shows long tails, but with the right tail noticeably longer than the left tail. We should consider an asymmetric error distribution.
F: We should not interpret (B) before testing for stationarity. First make an ADF test and, if the null hypothesis is not rejected, recalculate (B) when fitting to the differenced data.

Q4-03.



(A) Inflation (black) and unemployment (red) for the USA, 1980-2024. (B) Sample autocorrelation function of the residuals from a least square regression (`y=inflation, x=unemployment`), with estimated coefficients below. Which is the best interpretation of these graphs and fitted model?

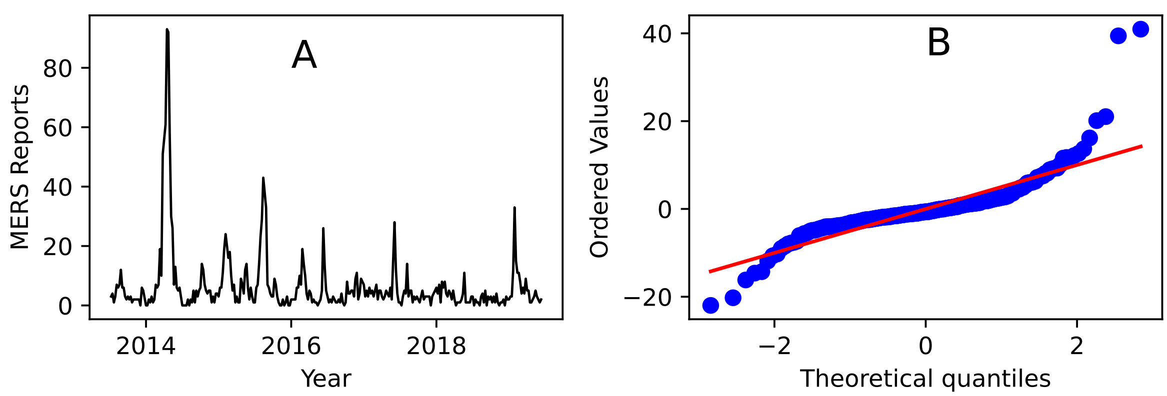
	coef	std err	t	P> |t|	[0.025	0.975]
const	2.8534	0.301	9.471	0.000	2.262	3.445
x1	0.0671	0.048	1.409	0.160	-0.026	0.161

A: 0.05 is a reasonable estimate for the additional unemployment caused by one percentage point of additional inflation. We should not trust the uncertainty estimate (not shown), since our model does not allow for autocorrelation of the residuals.

B: 0.05 is a reasonable estimate for the association between inflation and unemployment. We should not assume there is a causal relationship. We should not trust the uncertainty estimate (not shown), since our model does not allow for autocorrelation of the residuals.

C: 0.05 is a reasonable estimate for the association between inflation and unemployment. We should make an additional assumption that there are no confounding variables, and then we can interpret this association to be causal. We should not trust the uncertainty estimate (not shown), since our model does not allow for autocorrelation of the residuals.

Q4-04.



(A) Weekly cases of Middle East Respiratory Syndome (MERS) in Saudi Arabia. (B) a normal quantile residual plot for ARMA(2,2). We can formally test for non-normality of these residuals by a Shapiro-Wilk test (p-value=4.8×10^{-21}). What best describes the value added by presenting the Shapiro-Wilk test here?

A. We should always be alert for the danger of seeing patterns in noise. The Shapiro-Wilk test is useful to confirm our assessment that the normal quantile plot shows long tails.
B. Presenting the Shapiro-Wilk test here is not very insightful here, since the long tails are obvious from the normal quantile plot. However, adding this test demonstrates technical competence so it is better to include it than to omit it.

C. The long tails are established from the normal quantile plot. We could consider a log transform, or a long-tailed model, or a bootstrap simulation study to investigate whether the conclusions are sensitive to non-normality. Adding a fairly uninformative test instead of investigating the consequences of the error distribution could be a distraction from good data analysis.
D. The Shapiro-Wilk test is useful, but has the problem that it only tells us about lack of normality, not whether the non-normality is due to skew or kurtosis. We should supplement with a Jarque-Bera test to assess those.

Q4-05.

	AIC MA0	AIC MA1	AIC MA2	AIC MA3	AIC MA4	LBT MA0	LBT MA1	LBT MA2	LBT MA3	LBT MA4
AR0	174.82	48.81	9.61	-14.62	-17.98	7.6e-63	2.84e-26	2.28e-11	0.002	0.0409
AR1	-33.05	-34.44	-32.68	-30.69	-30.31	0.567	0.95	0.925	0.918	0.999
AR2	-34.07	-32.84	-33.25	-29.27	-28.51	0.951	0.926	0.98	0.954	0.996
AR3	-32.67	-33.2	-29.21	-29.91	-27.38	0.922	0.985	0.967	0.981	0.95
AR4	-30.77	-31.36	-30.11	-28.94	-24.92	0.92	0.967	0.989	0.978	0.994

The Ljung-Box test (LBT) provides an alternative approach to comparison of AIC values for selecting ARMA models. Whereas the standard sample autocorrelation function (ACF) residual plot tests each ACF component $\hat{\rho}_k$ under a null hypothesis of white noise, LBT tests $\sum_{k=1}^{h} \hat{\rho}_k^2$. Here, we present an AIC table and an LBT table (for $h = 5$). This course have favored AIC, with visual inspection of ACF and checking whether residual patterns appear in the frequency domain. There may be reasons to prefer LBT. Which of the following are good reasons to use LBT?

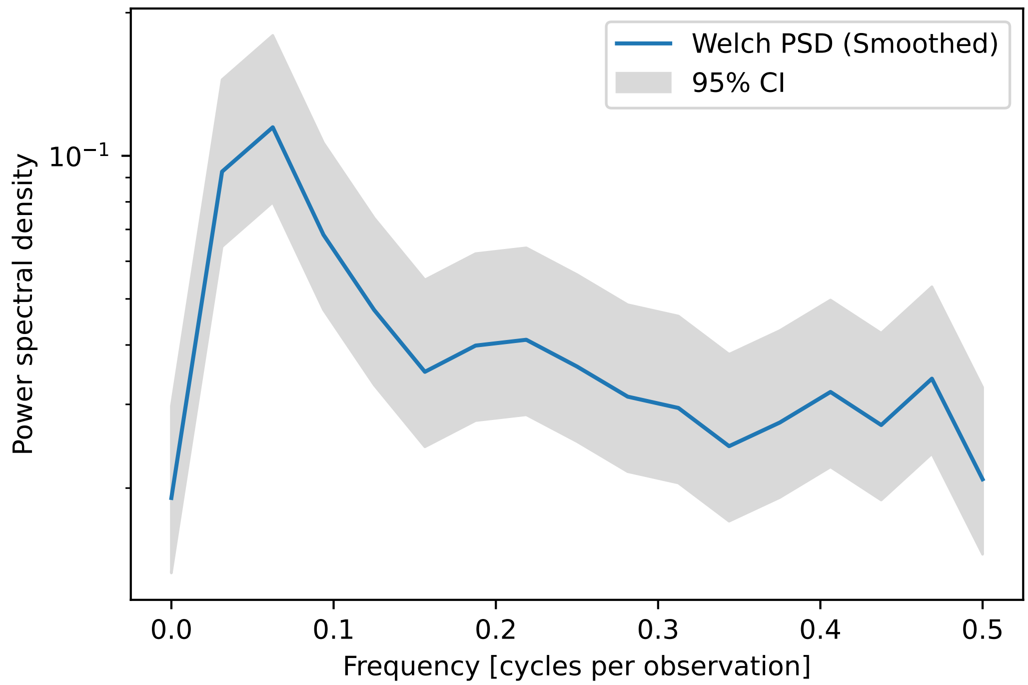
(i). LBT provides a p-value which is more formal than the comparison of AIC values.
(ii). Numerical issues involved in fitting an ARMA model may cause problems for comparing AIC values.
(iii). The LBT gives insights into what model to investigate next if the null hypothesis is rejected.
(iv). The LBT is useful in conjunction with AIC and ACF, since it provides an alternative perspective.

A. (i) only
B. (i, ii, iv)
C. (i,iii, iv)
D. (ii, iii, iv)
E. None of the above

Q5. The frequency domain

Q5-01.

We consider data $y_{1:415}$ where y_n is the time interval, in milliseconds, between the nth and $(n+1)$th firing event for a monkey neuron. Let $z_n = \log(y_n)$, with log being the natural logarithm. A smoothed periodogram of $z_{1:415}$ is shown below. Units of frequency are the default value in R, i.e., cycles per unit observation. We see a peak at a frequency of approximately 0.07.



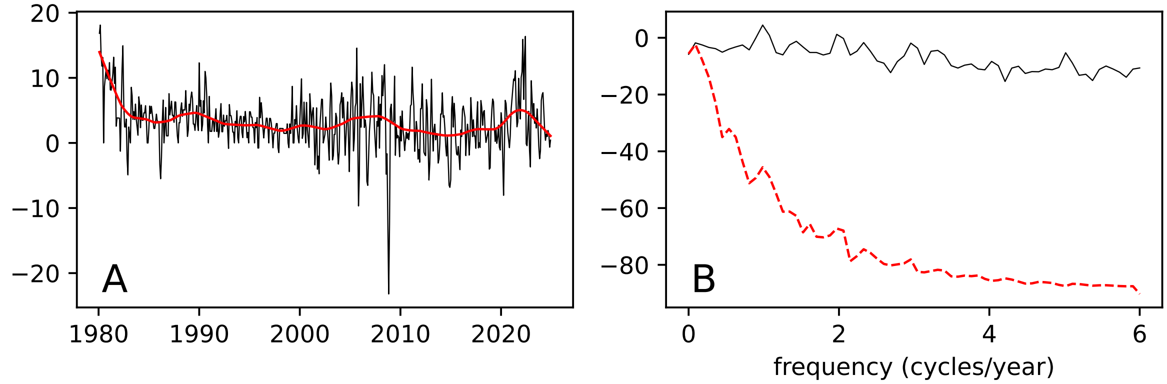
Which if the following is the best inference from this figure

A. Transitions between rapid neuron firing (short intervals between firing) and slow neuron firing (long intervals between firing) occur every $1/0.07 \approx 14$ firing events.

B. The neuron has a characteristic duration between firing events of $1/0.07 \approx 14$ milliseconds.

C. The neuron has a characteristic duration between firing events of $1/\exp(0.07) \approx 0.9$ milliseconds.

Q5-02.



The monthly US consumer price index (CPI) combines the price of a basket of products, such as eggs and bread and gasoline. (A) Annualized monthly percent inflation, i.e., the difference of log-CPI multiplied by 12×100 (black line); a smooth estimate via local linear regression (red line). (B) The periodogram of inflation and its smooth estimate. Which best characterizes the behavior of the smoother?

A: Cycles longer than 2 months are removed
B: Cycles shorter than 2 months are removed
C: Cycles longer than 2 year are removed
D: Cycles shorter than 2 year are removed
E: Cycles longer than (1/2) year are removed
F: Cycles shorter than (1/2) year are removed

Q6. Scholarship for time series projects

Q6-01.

This question on citing references applies to any statistics report, but it is particularly relevant here since we are learning proper use of sources in order to write open-access midterm and final projects.

Suppose that the midterm project P1 cites a past project, P2, in the reference list. P1 refererences P2 at one point, mentioning that the projects have similarities. When you look at the source code and the writing, you find various points where P1 and P2 are almost identical, though at other points the projects are entirely different. What do you infer?

A: The authors of P1 have done enough to honestly disclose the relationship with P2. After all, there is sufficient information provided for any reader to track down the exact relationship.

B: The authors of P1 have misrepresented the relationship with P2 by appearing to take credit for some original work which was in fact heavily dependent on a source. This is a serious offence which should be reported to Rackham and/or the Associate Chair for Graduate Programs in Statistics as a violation of academic integrity.

C: There is not enough information to tell the actual story for certain. The authors of P1 may or may not have done something wrong, depending on information that is not available to us, but they did cite P2 so they should be given the benefit of the doubt and should not lose any scholarship points.

D: The authors of P1 have misrepresented the relationship with P2 by appearing to take credit for some original work which was in fact heavily dependent on a source. This is a moderately severe offence, partly offset by including P2 in the reference list. A substantial number of scholarship points should be subtracted.

E: P1 evidently has not shown perfect scholarship, but this is a small issue that could easily be an honest mistake given that the authors were not trying to hide the fact that they had studied P2. It is appropriate to substract, say, 1 point for scholarship for this mistake.

Q6-02.

Four people in a team collaborate on a project. After the project is submitted, a reader identifies that part of the project is adapted from an unreferenced source, i.e., it has been plagiarized. The team worked using git and cooperates on tracking down the issue, and the commit history clearly reveals who wrote the problematic part of the project. What is the most appropriate course of action:

A. The guilty coauthor should be penalized heavily for poor scholarship, and the other coauthors should have a minor penalty for failing to check their colleague's work.

B. All coauthors should share the same penalty, since this is a team project and all coauthors share equal responsibility for the submitted report.

C. The guilty coauthor should be penalized heavily for poor scholarship. The other coauthors have demonstrated strong scholarship by following good transparent working practices that enabled this issue to get quickly resolved, so they should not receive any penalty.

D. It is necessary to collect more information before coming to a decision. For example, the team may argue that the source is well known to all readers so did not have to be cited.

Q6-03.

You discover that your team-mate is using Google Translate to carry out their share of the writing. The translation looks poorly done, similar in quality to ChatGPT, and does not use technical time series terminology correctly. What is the best course of action among the options below

A. Alert the instructor that you have a team mate adopting questionable scholarship strategies, in order to make sure you are not personally held responsible.

B. Ask ChatGPT to rewrite this problematic section to improve its quality

C. Help your team mate to rewrite the section in their own voice (shared with your voice).

Q6-04.

Why is it helpful for a course such as DATASCI/STATS 531, that permits the use of internet resources including GenAI and past solutions, to require students to say explicitly say when they do not use sources?

A. Failure to give credit to sources is against the academic integrity rules of Rackham, the graduate school at University of Michigan.

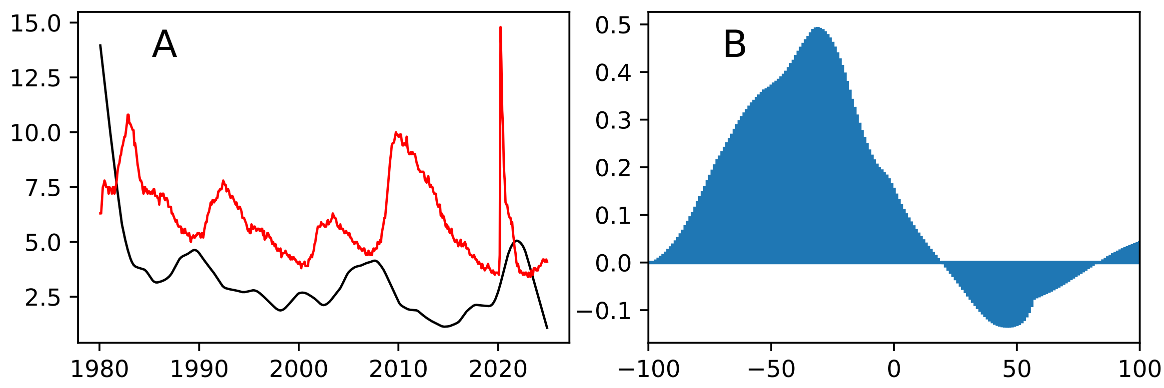
B. It helps the GSI to grade the homework when they know exactly what sources have been used and for what question.

C. Students whose solution is more dependent on sources than they want to admit are reluctant to explicitly deny using sources.

D. The GSI has the task of evaluating whether the student has demonstrated thought about the homework task beyond collecting material from sources into a solution. This is not an easy task even when the sources are clearly listed and referenced at the point (or points) where they are used.

Q7. Data analysis

Q7-01.



(A) Inflation (black) and unemployment (red) for the USA, 1980-2024. (B) Cross-correlation function, `plt.xcorr(inflation,unemployment)`. What is the best interpretation of this plot?

A: High inflation generally led high unemployment, with a lag of about 4 yr.
B: High inflation generally followed high unemployment, with a lag of about 4 yr.
C: Association is not causation, so we should not interpret a cross-correlation plot in terms of lead and lag relationships.

Q7-02.

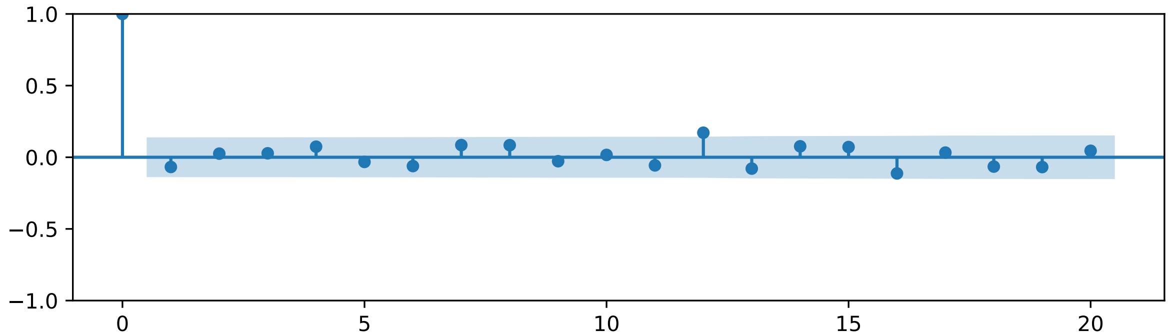
Which of the below best explains the role of models in time series analysis (and statistics more broadly)? Pick one, and explain your choice.

A. A statistical inference requires a statistical model. Investigating a hypothesis via a p-value requires a model to obtain the distribution of the p-value. Investigation of Bayesian posterior probabilities requires a model.

B. Parametric statistical tests require a model. However, where possible, we should use non-parametric methods (e.g., rank tests, or cross-validation) that do not require a model and so are more robust.

C. Modeling is of intrinsic scientific value, as a way of understanding the data-generating mechanism. We do not have to have specific statistical tests in mind when developing a model.

Q7-03.



The plot above is the sample autocorrelation function (ACF) for a time series $y_{1:N}$. What is the best conclusion to draw from this evidence?

A. The sample ACF is statistically consistent with an iid model. Therefore, standard statistical reasoning lets us conclude that the time series is iid.

B. The time series passed this test for being iid, but it might fail other tests. In practice, we can never make all possible tests so we cannot reliably conclude that the time series is iid.

C. It is meaningless to say that a time series is iid, since a time series is a sequence of numbers and iid is a property of a sequence of random variables.

D. The statistical evidence in the sample ACF is consistent with using an iid model to describe the data.

E. There is an oscillating pattern in the sample ACF. Even though no individual lag contradicts an iid model assumption at the 5% level, the overall pattern is clear evidence against iid.
