STATS 531 Homework 1

Due Sunday 1/18, 11:59pm

Submit your solution as a pdf via Canvas. You may scan a handwritten report. Later in the course, we will be using qmd for projects, so you might like to use that also for this homework. In that case, the source code for this assignment is available on GitHub to help get you started. Qmd (quarto) combines Python with Latex, so extra work will be required initially if you are unfamiliar with either of these.

The grading scheme for the homework report puts an emphasis on careful explanation of sources, as explained in the grading rubric. You are advised to read this rubric before submitting your homework. You are expected to support your homework report with sources, when appropriate. Usually, the sources are listed at the end of the report and cited where applicable within the report.

Questions about the homework can be asked and answered as GitHub issues on the course website. You can also directly contact the GSI (aaronabk@umich.edu) or instructor (ionides@umich.edu) but please post to the course website whenever that is appropriate.

Question 1.1. The covariance is $\mathrm{Cov}(X,Y) = E\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big]$. Recall the basic properties, following the convention that upper case letters are random variables and lower case letters are constants:

P1. $\mathrm{Cov}(Y,Y) = \mathrm{Var}(Y)$,

P2. $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$,

P3. $\mathrm{Cov}(aX,bY) = ab\,\mathrm{Cov}(X,Y)$,

P4. $\mathrm{Cov}\left(\sum_{m=1}^{M} Y_m, \sum_{n=1}^{N} Y_n\right) = \sum_{m=1}^{M}\sum_{n=1}^{N}\mathrm{Cov}(Y_m,Y_n)$.

Let $Y_{1:N}$ be a covariance stationary time series model with autocovariance function γ_h and constant mean function, $\mu_n = \mu$. Consider the sample mean as an estimator of μ,

$$\hat{\mu}(y_{1:N}) = \frac{1}{N}\sum_{n=1}^{N} y_n.$$

Show how the basic properties of covariance can be used to derive the expression,

$$\mathrm{Var}(\hat{\mu}(Y_{1:N})) = \frac{1}{N}\gamma_0 + \frac{2}{N^2}\sum_{h=1}^{N-1}(N-h)\gamma_h.$$

Question 1.2. The sample autocorrelation is perhaps the second most common type of plot in time series analysis, after simply plotting the data. We investigate how Python represents chance variation by the shaded region in the plot of the sample autocorrelation function produced by the `plot_acf` function in `statsmodels.graphics.tsaplots`. What approximation is being made? How should the region be interpreted statistically? Looking at the documentation is a good starting point:

```
import statsmodels as sm
help(sm.graphics.tsaplots.plot_acf)
```

From this help document, we read

```
Confidence intervals for ACF values are generally placed at 2
standard errors around r_k. The formula used for standard error
depends upon the situation. If the autocorrelations are being used
to test for randomness of residuals as part of the ARIMA routine,
the standard errors are determined assuming the residuals are white
noise. The approximate formula for any lag is that standard error
of each r_k = 1/sqrt(N)
```

However, checking exactly what was done requires looking at the actual code:

```
import inspect
print(inspect.getsource(sm.graphics.tsaplots.plot_acf))
```

It may be simpler to inspect the source code directly on GitHub, at https://github.com/statsmodels/statsmodels/blob/main/statsmodels/graphics/tsaplots.py. It appears that the plotted interval is determined by a quantity called `confint`, but some hunting is required to find where `confint` is determined. Eventually, looking at

```
print(inspect.getsource(sm.tsa.stattools.acf))
```

we find

```
    if bartlett_confint:
        varacf = np.ones_like(acf) / nobs
        varacf[0] = 0
        varacf[1] = 1.0 / nobs
        varacf[2:] *= 1 + 2 * np.cumsum(acf[1:-1] ** 2)
    else:
        varacf = 1.0 / len(x)
    interval = stats.norm.ppf(1 - _alpha / 2.0) * np.sqrt(varacf)
    confint = np.array(lzip(acf - interval, acf + interval))
```

For assessing the ACF of ARMA residuals, the help document suggests taking `barlett_confint=false`, corresponding to a normal distribution approximation for the sample autocorrelation, with mean zero and standard deviation $1/\sqrt{N}$.

A. This question investigates the use of $1/\sqrt{N}$ as an approximation to the standard deviation of the sample autocorrelation estimator under the null hypothesis that the time series is a sequence of independent, identically distributed (IID) mean zero random variables.

Instead of studying the full autocorrelation estimator, you are asked to analyze a simpler situation where we take advantage of the knowledge that the mean is zero and consider

$$\hat{\rho}_h(Y_{1:N}) = \frac{\frac{1}{N}\sum_{n=1}^{N-h} Y_n \, Y_{n+h}}{\frac{1}{N}\sum_{n=1}^{N} Y_n^2}$$

where $Y_1, \ldots, Y_N$ are IID random variables with zero mean and finite variance. Specifically, find the mean and standard deviation for $\hat{\rho}_h(Y_{1:N})$ when N becomes large.

The actual autocorrelation estimator subtracts a sample mean, and you can analyze that instead if you want an additional challenge.

You will probably want to make an argument based on linearization. You can reason at whatever level of math stat formalization you're happy with. According to *Mathematical Statistics and Data Analysis* by John Rice, a textbook used for the undergraduate upper level Math Stats course, STATS 426,

"When confronted with a nonlinear problem we cannot solve, we linearize. In probability and statistics, this method is called **propagation of errors** or the δ **method**. Linearization is carried out through a Taylor Series expansion."

Rice then proceeds to describe the delta method in a way very similar to the Wikipedia article on this topic. In summary, suppose X is a random variable with mean μ_X and small variance σ_X^2, and $g(x)$ is a nonlinear function with derivative $g'(x) = dg/dx$. To study the random variable $Y = g(X)$ we can make a Taylor series approximation,

$$Y \approx g(\mu_X) + (X - \mu_X) g'(\mu_X).$$

This approximates Y as a linear function of X, so we have

1. $\mu_Y = \mathbb{E}[Y] \approx g(\mu_X)$.

2. $\sigma_Y^2 = \mathrm{Var}(Y) \approx \sigma_X^2 \{g'(\mu_X)\}^2$.

3. If $X \sim N[\mu_X, \sigma_X^2]$, then Y approximately follows a $N[g(\mu_X), \sigma_X^2 \{g'(\mu_X)\}^2]$ distribution.

For this question, we have a two-dimensional situation, where $Y = g(U, V)$ and the Taylor series approximation becomes

$$Y \approx g(\mu_U, \mu_V) + (U - \mu_U)\frac{\partial}{\partial u} g(\mu_U, \mu_V) + (V - \mu_V)\frac{\partial}{\partial v} g(\mu_U, \mu_V)$$

with $U = \hat{\gamma}_h(Y_{1:N}) = \frac{1}{N}\sum_{n=1}^{N-h} Y_n \, Y_{n+h}$ and $V = \hat{\gamma}_0(Y_{1:N}) = \frac{1}{N}\sum_{n=1}^{N} Y_n{}^2$. Finding the mean and variance of U and V requires similar techniques to Question 1.1.

B. It is often asserted that the horizontal dashed lines on the sample ACF plot represent a confidence interval. In particular, this is done in the help documentation for `sm.graphics.tsaplots.plot_acf`, copied above. Use a definition of a confidence interval to explain how the shaded interval in the resulting plot does, or does not, construct a confidence interval.