BOSTON UNIVERSITY

Term Paper - the following will walk students through an example of applying some of the mathematical methods we have learned thus far. It should be considered a heuristic or guide to fulfilling this requirement for the course.

Let's apply what we've learned thus far in terms of mathematical modeling. Using a linear model to explain the evolution of the returns of a stock, we'll perform some tests, by predicting "future" outcomes.

PART 1

1) Select an equity instrument trading in the US stock market and download its closing daily price for a 3-year period. It's probably better to find more recent data (e.g., 2020-23). Note the name of the security, identifying ID, etc. Also note whether there are dividends paid.

2) We need a benchmark. Based upon the instrument/security you select, find an appropriate benchmark for your security.

Large-cap	$10 billion - $200 billion
Mid-cap	$2 billion - $10 billion
Small-cap	$250 million - $2 billion
Micro-cap	less than $250 million
Nano-cap	less than $50 million

   For example, if you select a security that falls in the range of a mid-cap security, you can select a benchmark of the Standard & Poors 400. Ticker symbol `MDY`.

3) Daily Returns. Most of the estimations will refer to returns rather than prices (although we'll work with the prices at the end), you need to create a return series both for your stock, for your benchmark.

   There are two types of returns you can use, logarithmic and arithmetic. I prefer logarithmic returns so please construct those.
   - Between dates t and t+1, and using the original prices P (levels), you can construct returns as follows:
   - A logarithmic return is defined for day t+1 is $ln(p(t+1)/p(t))$.
   - An arithmetic return is defined for t+1 as $(p(t+1) - p(t)) / p(t)$.
   - Thus, you can construct returns from day 2 to day T (last). The returns for day 1 would need the price at time 0.

4) Create a table where you report the
   - Name
   - Ticker
   - Mean
   - Median
   - Sample standard deviation
   - Minimum
   - Maximum
   - Sample skewness
   - Sample kurtosis
   - Starting date
   - Final date

   for your stock returns

5) Separate the data. Create two separate data periods
   - First 2 years. This data will be used to "teach" or establish our model.
   - Last year. This data will be used to test our model and to "forecast".
   - Thus 2 tables (for the 2 different periods) for each of the 3 series.
   - Indicate if there are noticeable differences (or not) between the summary stats in the second period relative to the first period.

6) Estimate an OLS regression for your stock returns and intercept. Repeat that exercise with your benchmark returns. Also estimate it with the Newey-West (1 lag) correction to account for potential AC and HET. See Section 16.4 on "*Heteroskedasticity and Autocorrelation Consistent Standard Errors*" for an additional reference.

   - For this and all regressions below, use only the first two (2) years of your sample.
   - Leave the last year of your sample untouched so later you can do out-of-sample forecasts.

7) DECIDE on a 2nd regressor variable X_2 for your stock and get the data for the same sample as your stock. X_2 can NOT be another stock. X_2 may be another index, an input price return (oil, gold, etc.) or a macro variable (inflation, interest rates, economic activity, etc.).
   If X_2 is in levels, convert it to log returns. e.g., if you got gold prices, convert them to gold log returns. See step 3.

8) Estimate an OLS regression for your stock returns vs intercept, benchmark returns, and X_2 returns. Also estimate it with the Newey-West (1 lag) correction. Also, estimate an OLS regression for your stock returns vs intercept, CRSP returns and X_2 returns. Also estimate it with the Newey-West (1 lag) correction.

9) Present the 4 estimated regressions as in the table below, ordered by column (1) to (4). Cells with an "x" indicate what cells should be filled for each regression.

Regression of ABC LogReturns				
Sample: January 2, 2018 to December 31, 2019				
	1	2	3	4
Intercept	x	x	x	x
(OLS s.e.)	x	x	x	x
(NW s.e.)	x	x	x	x
S&P500	x		x	
(OLS s.e.)	x		x	
(NW s.e.)	x		x	
CRSP		x		x
(OLS s.e.)		x		x
(NW s.e.)		x		x
X2			x	x
(OLS s.e.)			x	x
(NW s.e.)			x	x
R2 Adjusted	x	x	x	x
s.e. Reg	x	x	x	x
NOBS	x	x	x	x

- There's one row representing each of the variables (intercept, benchmark, security and X_2). In this row, write the estimated coefficient (beta hat) for that variable in the respective regression column.

- The line below corresponds to the OLS standard errors of each of the estimated coefficients, which should be put in parentheses below these coefficients.

- The line below corresponds to the Newey-West(1) standard errors of the estimated coefficients, which should be put in square parentheses below the OLS standard errors.

- Also present the R^2 Adjusted, the SE regression and the number of observations for each of the regressions (columns).

10) Indicate for each of the 4 regressions which variables are significant and which are not, both using OLS t-statistics and Newey-West t-statistics.

    Does any variable change significance when you compare the OLS t-statistic vs the Newey-West t-statistic? If so, what is the explanation? If not, why?

11) Compare regressions (1) and (2). Is there any evidence of omitted variable bias? Explain briefly.

12) Compare regressions (3) and (4) based on all you know from this course. Which one would you prefer to use as the main specification for your stock regression? Why?

Appendix – Newey-West and Heteroskedasticity and Autocorrelation Consistent Standard Errors (HAC)

The Newey-West standard errors, also known as heteroskedasticity and autocorrelation consistent (HAC) standard errors, are used in econometrics to provide consistent estimates of the standard errors of regression coefficients in the presence of heteroskedasticity and autocorrelation of unknown form.

In the presence of heteroskedasticity and autocorrelation, which are common in econometric analyses, particularly with time series data. By adjusting the standard errors, researchers can avoid misleading conclusions that could arise from incorrect standard error estimates.

- Heteroskedasticity - This refers to the situation where the variance of the errors in a regression model is not constant across observations. Ordinary Least Squares (OLS) standard errors can be biased in the presence of heteroskedasticity, leading to incorrect statistical inferences.

- Autocorrelation - This occurs when the residuals (errors) from a regression model are correlated across time. This is common in time series data and can also lead to biased standard error estimates.

Some historical context regarding the Newey-West Standard Errors.
- Developed by Whitney K. Newey and Kenneth D. West in 1987.
- Provide a method to adjust the standard errors of the OLS estimators to account for both heteroskedasticity and autocorrelation.
- The adjustment ensures that the standard errors are consistent, meaning that they converge to the true standard errors as the sample size increases, even if the errors are heteroskedastic or autocorrelated.

Here's a simplified outline of the process. The Newey-West standard errors are computed using a weighted sum of the auto-covariances of the residuals.

- Fit the OLS Model. Estimate the regression model using ordinary least squares to obtain the residuals

- Compute Residuals. Calculate the residuals from the fitted OLS model

- Calculate Auto-covariances. Compute the autocovariances of the residuals for different lags up to a specified maximum lag length q

- Apply Weights. Apply weights to the autocovariances. These weights typically decrease linearly with the lag, reflecting the assumption that autocorrelation diminishes with distance

- Form the Variance-Covariance Matrix. Use the weighted autocovariances to form the Newey-West estimator of the variance-covariance matrix of the regression coefficients.

Formula. The Newey-West estimator for the variance-covariance matrix of the OLS estimator $\hat{\beta}$ is given by

$$\hat{V}_{NW} = \sum_{k=-q}^{q} w(k,q)\hat{\Gamma}(k)$$

where:

- $\hat{\Gamma}(k)$ is the sample autocovariance at lag k.
- $w(k,q)$ is the weight applied to the autocovariance at lag k. A common choice is the Bartlett kernel weight, which is $w(k,q) = 1 - (|k| / q+1)$.

Some more practical points in terms of using Newey-West.

- Software Implementation. Many statistical software packages (e.g., R, Stata, Python) have built-in functions to compute Newey-West standard errors. For instance, in R, the sandwich package provides functions like `vcovHAC()` for this purpose. In Python, it's `sw.cov_hac()` where `sw` is the `statsmodel` library.

- Choice of Lag Length *q.* The choice of the maximum lag length q is crucial. A common rule of thumb is to set q to a function of the sample size n, such as $q = 4(n/100)^{2/9}$.