**Advanced Financial Statistics 2020/2021**

**Marked Assignment**

**Due 11 March 2021**

**1.** (This problem is based on Empirical Exercise E15.1 in Stock, J. and Watson, M., *Introduction to* Econometrics, 4th Edition. Please feel free to use this and other references when answering the questions in the assignment).

The file **USMacro_Quarterly** contains quarterly data on several macroeconomic time series for the United States. The variable *PCED* is the price index for personal consumption expenditures from the U.S. National Income and Product Accounts (the Personal Consumption Expenditures Deflator). In this exercise, you will construct forecasting models for the rate of inflation based on *PCED.* For this analysis, use the sample period 1963:Q1-2017:Q4 (where data before 1963 may be used, as necessary, as initial values for lags in regressions).

**a.** i. Compute the inflation rate $Inf = 400 \times [\ln(PCED_t) - \ln(PCED_{t-1})]$. What are the units of *Inf*? (Is *Inf* measured in dollars, percentage points, percentage per quarter, percentage per year, or something else? Explain).

ii. Plot the value of *Inf* from 1963:Q1 through 2017:Q4. Based on the plot, do you think that *Inf* has a stochastic trend? Explain.

**b.** i. Compute the first four autocorrelations of $\Delta Inf$.
ii. Plot the value of $\Delta Inf$ from 1963:Q1 through 2017:Q4. Explain why the behaviour of $\Delta Inf$ in the plot is consistent with the first order autocorrelation that you computed in (i).

**c.** i. Run an OLS regression of $\Delta Inf_t$ on $\Delta Inf_{t-1}$. Does knowing the change in inflation over the current quarter help predict the change in inflation over the next quarter? Explain.

ii. Estimate an AR(2) model for $\Delta Inf$. Is the AR(2) model better than an AR(1) model? Explain.

iii. Estimate an AR(*p*) model for $p$=0, 1, …, 8. What lag length is chosen by the BIC? What lag length is chosen by the AIC?

iv. Use the AR(2) model to predict the change in inflation from 2017:Q4 to 2018:Q1 – that is, to predict the value of $\Delta Inf_{2018:Q1}$.

v. Use the AR(2) model to predict the level of the inflation rate in 2018:Q1 – that is, $Inf_{2018:Q1}$.

**d.** i. Use the ADF test for the regression

$$\Delta Inf_t = \beta_0 + \delta Inf_{t-1} + \gamma_1 \Delta Inf_{t-1} + \gamma_2 \Delta Inf_{t-2} + u_t$$

with two lags of $\Delta Inf$ to test for a stochastic trend in *Inf.*

ii. Is the ADF test based on equation (1) preferred to the test based on ADF regression with time trend, e.g., on the ADF regression

$$\Delta Inf_t = \beta_0 + \alpha t + \delta Inf_{t-1} + \gamma_1 \Delta Inf_{t-1} + \gamma_2 \Delta Inf_{t-2} + u_t?$$

Explain.

iii. In (i), you used two lags of $\Delta Inf$. Should you used more lags? Fewer lags? Explain.

iv. Based on the test you carried out in (i), does the AR model for *Inf* contain a unit root? Explain carefully. (*Hint*: Does the failure to reject a null hypothesis mean that the null hypothesis is true?)

**e.** Use the QLR test with 15% trimming to test the stability of the coefficients in the AR(2) model for $\Delta Inf$. Is the AR(2) model stable? Explain.

**2.** In the data file **USMacro_Monthly**, you will find data on two aggregate price series for the United States: the Consumer Price Index (CPI) and the Personal Consumption Expenditures Deflator (PCED). These series are alternative measures of consumer prices in the United States. The CPI prices a basket of goods whose composition is updated every 5-10 years. The PCED uses chain-weighting to price a basket of goods whose composition changes from month to month. Economists have argued that CPI will overstate inflation because it does not take into account the substitution that occurs when relative prices change. If this substitution bias is important, then average CPI inflation should be systematically higher than that PCED inflation. Let $\pi_t^{CPI} = 1200 \times \ln[CPI(t)/CPI(t-1)]$, $\pi_t^{PCED} = 1200 \times \ln[PCED(t)/PCED(t-1)]$, and $Y_t = \pi_t^{CPI} - \pi^{PCED}$, so that $\pi_t^{CPI}$ is the monthly rate of price inflation (measured in percentage points at an annual rate) based on the CPI, $\pi_t^{PCED}$ is the monthly rate of price inflation from the PCED, and $Y_t$ is the difference. Using data from 1959:1 through 2004:12, carry out the following exercises.

Consider the "constant-term-only" regression

$$Y_t = \mu + u_t,\ t=1,\ \ldots,\ T. \tag{1}$$

Do you think $u_t$ is serially correlated? Explain.

*Inference on $\mu$ in model (1) under heteroskedasticity and autocorrelation in $u_t$ is usually based on the t-statistic $t_{\bar{Y}} = \dfrac{\sqrt{T}(\bar{Y}-\mu_0)}{\hat{\omega}}$, where $\mu_0$ is a hypothesized value of $\mu$, $\bar{Y} = \dfrac{1}{T}\sum_{t=1}^{T} Y_t$ is the sample mean and $\hat{\omega}^2$ is an estimator of the long-run variance $\omega^2 = \lim_{T\to\infty} T\ Var(\bar{Y})$ of $Y_t$.*

**(a)** Construct a 95% confidence interval for $\mu$ using Newey-West HAC standard errors with $\omega^2$ estimated by

$$\hat{\omega}^2(b) = \hat{\gamma}(0) + 2\sum_{h=1}^{T-1} k(h/b)\hat{\gamma}(h), \tag{2}$$

where $\hat{\gamma}(h)$ are sample autocovariances, $k(z)$ is the Bartlett kernel satisfying $k(z) = 1 - z$ for $0 \le z \le 1$, $k(z) = 0$ for $z > 1$, and *b=b(T)* is the HAC bandwidth parameter. What value of the HAC bandwidth parameter *b* did you chose? Why? Is there statistically significant evidence that the mean inflation rate for the CPI is greater than the rate for the PCED?

**(b)** State the assumptions that guarantee consistency of the standard errors you used in part (b) and discuss whether they are likely to hold in this exercise.

## References

Stock and Watson, Ch. 15.
Hamilton (1994), Ch. 10
Newey, W. and West, K. (1987). A simple positive semi-definite, heteroskedastic and autocorrelation consistent covariance matrix. *Econometrica* **55**, 703-708.
Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* **59**, 817-858.

**(c)** Let $X_t$, $t$=1, …, $q$, $q \geq 2$, be independent mean-zero normal $N(\mu, \sigma_t^2)$ random variables with common mean $E(X_t) = \mu$ and variances $Var(X_t) = \sigma_t^2$. Consider the usual $t$-statistic for the hypothesis test $H_0 : \mu = 0$ against $H_a : \mu > 0$ given by $t(X_1, ..., X_q) = t_{\overline{X}} = \sqrt{q}\frac{\overline{X}}{s_X}$, where $\overline{X} = \frac{1}{q}\sum_{t=1}^{q} X_t$ is the sample mean and $s_X^2 = \frac{1}{q-1}\sum_{i=1}^{q}(X_i - \overline{X})^2$ is the sample variance of $X_t$, $t$=1, …, $q$. As discussed in Ibragimov and Müller (2007) (see Theorem 1 in that paper and the discussion following it), the one-sided $t$-test using $t_{\overline{X}}$ is conservative in the following sense:

Let $T_{q-1}$ denote a random variable that has Student-$t$ distribution with $q-1$ degrees of freedom. Further, let $cv_q(\alpha)$ be the critical value of the usual one-sided $t$-test based on $t_{\overline{X}}$ of level $\alpha$, that is, $P(T_{q-1} > cv_q(\alpha)) = \alpha$. Then, if $2 \leq q \leq 14$ and $\alpha \leq 0.05$, then

$$\sup_{\sigma_1^2, ..., \sigma_q^2} P(t_{\overline{X}} > cv_q(\alpha)) = P(T_{q-1} > cv_q(\alpha)) = \alpha . \tag{3}$$

Further, relation (2) holds for all $q \geq 2$ if $\alpha \leq \Phi(-\sqrt{3}) = 0.04163$.

The $t$-statistic based approach to asymptotic inference in Ibragimov and Müller (2010) can be thus used to test $H_0 : \mu = 0$ against $H_a : \mu > 0$ in model (1). One first divides the data into $q$ blocks and estimates the sample means $\overline{Y}^{(s)}$ using the data in each of the blocks $s$=1, …, $q$. Then the statistic $t(\overline{Y}^{(1)}, ..., \overline{Y}^{(q)})$ is formed using the estimators and the null hypothesis is rejected in favor of the alternative for large positive values of $t(\overline{Y}^{(1)}, ..., \overline{Y}^{(q)})$ (see Ibragimov and Müller (2010) for further discussion). Relation (3) implies that the procedure results in asymptotically correct inference as long as the significance level $\alpha \leq \Phi(-\sqrt{3}) = 0.04163$ for $q \geq 14$ and $\alpha \leq 0.05$ for $2 \leq q \leq 14$ and, under $H_0$, $(\sqrt{T}\,\overline{Y}^{(1)}, ..., \sqrt{T}\,\overline{Y}^{(q)}) \xrightarrow{d} (Z_1, ..., Z_q)$, where $Z_i$ , $i$=1, …, $q$, are independent normal $N(0, \omega_i^2)$ random variables with possibly different limit variances $\omega_i^2$. The approach is thus robust to, for instance, changes in the long-run variance of $Y_t$. By conditioning, it can also be used in the case of convergence to mixtures of normals with stochastic $\omega_i^2$ which is important, for instance, in the analysis of heavy-tailed models and stochastic volatility and conditional heteroskedasticity models.

Apply the *t*-statistic based approach described to test $H_0 : \mu = 0$ against $H_a : \mu > 0$ in model (1) and compare your results to those in part (a).

**(d)** Discuss your choices for the number of blocks *q* in part (c). Discuss whether the assumptions for asymptotic validity of the *t*-statistic based procedure are likely to be satisfied in the exercise.

## Reference

Ibragimov, R. and Müller, U. (2010). *t*-statistic based correlation and heterogeneity robust inference. *Journal of Business and Economic Statistics* 28, 453-468.

**3.** Recall that, according to empirical results we discussed in class, (daily) financial returns ($r_t$) in developed countries typically have tail indices $\zeta$ in the interval (2, 4). That is, the tails of distributions of these variables follow a power law: $P(|r_t| > x) \sim \frac{C}{x^{\zeta}}$ for large $x>0$, with $\zeta \in (2, 4)$. This implies, in particular, that the variances of the financial returns in developed economies are finite.

**(a)** Using a financial return time series of your choice over the time period that contains the beginning of the 2008 financial crisis (Sept. 2008), estimate its tail index $\zeta$ using Hill's and log-log rank-size regression approaches.

**(b)** Construct the corresponding confidence intervals for the true value of the tail index $\zeta$ using its Hill's and log-log rank-size estimates.

**(c)** Do the obtained tail index estimates and confidence intervals support the conclusion that the tail index of the return time series you consider lies in the interval (2, 4)? Do the support finiteness of variances for the time series?

*Note*: In this problem, you can consider, for instance, a time series of daily returns on a developed or emerging country financial index like S&P 500, DJIA or FTSE; a time series of daily returns on a particular stock; or a time series of bitcoin or other cryptocurrency returns.

**References.**

Ibragimov, M., Ibragimov, R. and Walden, J. (2015). *Heavy-Tailedness and Robustness in Economics and Finance.* Lecture Notes in Statistics 214, Springer; Ch. 3.

Ibragimov, M., Ibragimov, R. and Kattuman, W. (2013). Emerging markets and heavy tails. *Journal of Banking and Finance* **37**, pp. 2546-2559.