Total assignment points **(100 points)**

1. **Regression Diagnostics Checking** **(45 points)**

   Follow the steps outlined in Appendix A and download the monthly stock price time series of Coca Cola from WRDS. Next, open Kenneth French's homepage and download the "Fama/French 5 factors (2x3)". The file contains the monthly time series of the five factors Fama and French (2015, 2016) propose in their two latest papers as an extension of their conventional three-factor model in which they suggest two new factors: operating profitability and investment.[1,2] Thus they consider the following five factors:

   - Mkt-Rf (R^e_M): the excess return on the market.
   - SMB (Small Minus Big): the average return on the *small* stock portfolios minus the average return on the *big* stock portfolios.
   - HML (High Minus Low): the average return on the *value* portfolios minus the average return on the *growth* portfolios.
   - RMW (Robust Minus Weak): the average return on the *robust* operating profitability portfolios minus the average return on the *weak* operating profitability portfolios.
   - CMA (Conservative Minus Aggressive): the average return on the *conservative* investment portfolios minus the average return on the *aggressive* investment portfolios.

[1] Details on how Fama and French (1992, 1993) constructed the factors SMB and HML can be found in Appendix B.

[2] The Fama-French five factors are constructed using six value-weighted portfolios formed on size and book-to-market, six value-weighted portfolios formed on size and operating profitability, and six value-weighted portfolios formed on size and investment.

Moreover, the last column contains the monthly time series for the simple return on the US-Treasury bill with one month to maturity (which we will use as proxy for the risk-free rate in this exercise). [3]

(a) First, calculate the *simple* return in *excess* of our proxy for the risk-free rate for Coca Cola. Report the (arithmetic) mean and the standard deviation for the resulting excess return time series and each of the five factors in your solution paper.

After that, use the five Fama-French factors at time t as the independent variables and the excess return of Coca Cola at time t as the dependent variable and run the following regression:

$$R^e_{CC,t} \quad = \quad \beta_1 + \beta_2 \, R^e_{M,t} + \beta_3 \, \mathrm{SMB}_t + \beta_4 \, \mathrm{HML}_t + \beta_5 \, \mathrm{RMW}_t + \beta_6 \, \mathrm{CMA}_t + u_{CC,t}.$$

Report the parameter estimates, their t-statistics, and the adjusted R^2 in your solution paper. Based on a 5% significance level, which of the variables are statistically significantly different from zero? Is there a variable you would consider deleting from the regression? Explain your answer. For each variable which is statistically significant, provide an assessment of its economic significance by looking at the impact of a one-standard deviation change in the corresponding variable. *(10 points)*

(b) For each of the following hypotheses, provide an economic interpretation and run the corresponding test using a significance level of 5%:

i. H_0: $\beta_2 = 0$ and $\beta_3 = 0$ and $\beta_4 = 0$ and $\beta_5 = 0$ and $\beta_6 = 0$ vs.
H_1: $\beta_2 \neq 0$ or $\beta_3 \neq 0$ or $\beta_4 \neq 0$ or $\beta_5 \neq 0$ or $\beta_6 \neq 0$ *(2 points)*

ii. H_0: $\beta_3 = 0$ and $\beta_4 = 0$ and $\beta_5 = 0$ and $\beta_6 = 0$ vs.
H_1: $\beta_3 \neq 0$ or $\beta_4 \neq 0$ or $\beta_5 \neq 0$ or $\beta_6 \neq 0$ *(2 points)*

iii. H_0: $\beta_5 = 0$ and $\beta_6 = 0$ vs.
H_1: $\beta_5 \neq 0$ or $\beta_6 \neq 0$ *(2 points)*

Next, check (step-by-step) whether the assumptions of the classical linear regression model (CLRM) hold.

(c) Start with the first CLRM assumption, $E[u_{CC,t}] = 0$. Test the null hypothesis that the mean of the residuals is zero against the two-sided alternative. Give the p-value of the corresponding test. Can you reject the null hypothesis at the 5% significance level? Is the first assumption of the classical linear regression model violated? *(2 points)*

(d) Next turn to the second CLRM assumption of homoscedasticity, i.e., $Var[u_{CC,t}] = \sigma^2 < \infty$. Run White's test. Give the p-value of the corresponding F- and χ^2-tests. Can you reject the null hypothesis against the two-sided alternative at the 5% significance level? Is the second assumption of the classical linear regression model violated? *(4 points)*

(e) Run the regression from above using White (1980)'s estimation correction which produces heteroscedasticity consistent standard errors. Give the parameter estimates, their t-statistics, and the adjusted R^2. Based on a 5% significance level which of the variables are statistically significantly different from zero? Compare the adjusted R^2 you just calculated to the one from your previous OLS regression. Did it remain unchanged? Explain why is this the case. *(7 points)*

[Hint: Follow the steps in Appendix C and use the `olsnw` function (which is part of the MFE toolbox) to run this regression.]

[3]Note that the risk-free rate and Fama-French factor returns are expressed in percentage points whereas the returns on Coca Cola are expressed in decimal points.

(f) The third assumption of the CLRM states that the covariance between the residuals over time is zero, i.e., $Cov(u_{CC,i}, u_{CC,j}) = 0$ for $i \neq j$. Run a Breusch-Godfrey test for up to tenth-order autocorrelation (i.e., with ten lags). Give the p-value of the corresponding F- and χ^2-tests. Can you reject the null hypothesis against the two-sided alternative hypothesis at the 5% significance level? Is the third assumption of the classical linear regression model violated? *(4 points)*

(g) The fourth assumption of the CLRM states that there is no relation between the residuals and the independent variables, i.e., $Cov(u_{CC,t},\, x_{i,t}) = 0$ for all $i = 2, \ldots, k$ where k is the number of independent variables (including a constant). Compute the correlation between the estimated residual and each of the five Fama-French factors. Is the fourth assumption of the classical linear regression model violated? *(2 points)*

(h) Next, turn to the fifth assumption that the residuals are normally distributed with mean zero and variance σ^2, i.e., $u_{CC,t} \sim N(0, \sigma^2)$. Give the p-value of the Jarque-Bera test. Can you reject the null hypothesis underlying the Jarque-Bera test against the two-sided alternative at the 5% significance level? Is the fifth assumption of the classical linear regression model violated? *(3 points)*

(i) Create a histogram and a time series plot of the estimated residuals ($\widehat{u_{CC,t}}$). Suppose there is a theoretical reason to introduce a separate dummy variable for each point in time when the estimated residual is above or below 13, and zero otherwise. If there are, for instance, three points in time, you will need three dummy variables. Run the new regression including the dummy variables. Give the p-value of the corresponding Jarque-Bera test. Is the fifth assumption of the classical linear regression model violated? Explain how the presence of the dummy variables affects the estimated residuals that were previously above or below 13.

[Please do not report the regression results in your solution paper, just include them in the output of your code in the appendix.] *(7 points)*

2. **Multivariate Regression Analysis** **(30 points)**

Start by entering the following command in MATLAB: `rng(10, 'twister')`. Next, create 100 normally distributed random variables with 100 observations each and refer to those as your independent variables $x_1, \ldots, x_{100}$. Create a vector with the same number of observations which represents your dependent variable y.

(a) Run the following regression *stepwise*:

$$y_t = \alpha + \beta_1\, x_{1,t} + \beta_2\, x_{2,t} + \ldots + \beta_{100}\, x_{100,t} + u_t.$$

Report the five *jointly* most "important" explanatory variables selected by the stepwise regression (which we will refer to as $z_1, \ldots, z_5$). Use those as the independent variables, the vector y as the dependent variable, and run the following regression:

$$y_t = \alpha + \beta_1\, z_{1,t} + \beta_2\, z_{2,t} + \ldots + \beta_5\, z_{5,t} + v_t.$$

Report the parameter estimates, their t-statistics, and the R^2 in your solution paper. Based on a 5% significance level, discuss the statistical significance and the economic significance of your results. *(10 points)*

(b) A friend of yours who has a PhD in Finance argues that your analysis is related to a recent empirical study by Chordia et al. (2017) which is available as an attachment to this assignment. Read the introduction of the paper. Explain what your friend is referring to and why she might be right. *(20 points)*

3. **Univariate Time Series Analysis** **(25 points)**

Suppose you want to write down a model for monthly (seasonally adjusted) time series of the industrial production index (ipi_t). Up to now, it is unclear how such a model should look like. In this exercise, you are supposed to come up with the most suitable model specification.

Download the above mentioned time series (ipi_t) at the monthly frequency from the Fred homepage for the longest time period that is available, January 1919 until December 2020. It contains the (seasonally adjusted) industrial production index from the Federal Reserve Bank of St. Louis. The seasonal adjustment removes seasonal patterns from the data.

(a) First of all, you need to decide on the most appropriate time series for modelling. Plot the following time series: $\log(\text{ipi}_t)$ and $\log\left(\frac{\text{ipi}_t}{\text{ipi}_{t-1}}\right)$. Which of the two is more suitable for modelling? Explain your answer. *(5 points)*

(b) Having established the time series, your next step is to find the most appropriate model specification. Consider the following five alternatives: AR(1), AR(4), MA(1), MA(4), and ARMA(4,3). For each of the models, report the parameter estimates, the corresponding t-statstics, and Schwarz's Bayesian information criterion. The latter is calculated for an ARMA(p,q) model as follows

$$\text{SBIC} = \ln\left(\hat{\sigma}^2\right) + \frac{k}{T}\ln(T),$$

where $\hat{\sigma}^2 = \frac{\sum_{t=1}^{T}\hat{u}_t^2}{T-k}$ is the estimated error variance, $k = p + q + 1$ is the total number of parameters to be estimated, and T is the number of observations. Which is the most suitable model specification? Explain your answer. *(20 points)*

References

Chordia, T., A. Goyal, and A. Saretto. 2017. p-hacking: Evidence from two millian trading strategies. Working Paper.

Fama, E., and K. French. 1992. The Cross-Section of Expected Stock Returns. *Journal of Finance* 47:427–465. doi:10.1111/j.1540-6261.1992.tb04398.x.

Fama, E., and K. French. 1993. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33:3–56. doi:10.1016/0304-405X(93)90023-5.

Fama, E., and K. French. 2015. A five-factor asset pricing model. *Journal of Financial Economics* 116:1–22. doi:10.1016/j.jfineco.2014.10.010.

Fama, E., and K. French. 2016. Dissecting anomalies with a five-factor model. *Review of Financial Studies* 29:69–103. doi:10.1093/rfs/hhv043.

Newey, W., and K. West. 1987. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* 55:703–708. doi:10.2307/1913610.

White, H. 1980. A Heteroskedasticity-Consistent Covariance Matrix and a Direct Test for Heteroskedasticity. *Econometrica* 48:817–838. doi:10.2307/1912934.

Appendix

A Downloading returns on Coca Cola

(1) Open the this homepage and log in to WRDS.[4]

(2) Scroll down and click on "I accept and agree to these terms and conditions" (of course, only if you really accept the WRDS terms of use).

(3) Once the "Thank you" page opens, hit "Home" which you can find in the upper left corner.

(4) Click on "CRSP", and choose "Stock / Security Files".

(5) Select "Monthly Stock File" and fill out the query form.

   (a) Specify the date range from "1944-02" to "2020-12" which is the longest time period that is available.

   (b) Next, you need to identify the stock you wish to download. Click on "Ticker" and use `KO` as company code.

   (c) After that, you have to select the variables you want to download. Choose `Holding Period Return`. These are simple stock returns adjusted for both dividend distributions and stock splits.

   (d) Finally, choose "Excel Spreadsheet" as output format, and click on "Submit query".

(6) A new window pops up, and after waiting for a (short) while, your Excel file is created which can be downloaded right away.

B Fama-French three factor portfolios

Fama and French (1992, 1993)'s starting point are all the firms in the CRSP database (i.e., all common stocks which are traded at the NYSE, AMEX, and NASDAQ). For each of those firms, the authors calculate the two quantities:

(1) market equity (also known as market capitalization), i.e., stock prices times number of shares

(2) book-to-market ratio, i.e., book equity relative to market equity.

Then Fama and French sort all the firms (independently) according to each of these quantities, from the lowest to the highest.

(1) The firms sorted with respect to market equity (also known as market capitalization) are divided into two portfolios. Stocks which have a market equity...

   - smaller than the median market equity of all NYSE traded stocks are considered as *small*.
   - larger or equal to the median market equity of all NYSE traded stocks are considered as *big*.

(2) The firms sorted with respect to the book-to-market ratio are divided into three portfolios. Stocks which have a book-to-market ratio ...

[4] You can find the username and password in the "WRDS" note in the "data" folder on itslearning.

- smaller than the 30th percentile of the book-to-market ratio of all NYSE traded stocks are considered as *growth* stocks.
- smaller than the 30th percentile **and** larger or equal to the 70th percentile of the book-to-market ratio of all NYSE traded stocks are considered as *neutral* stocks.
- larger or equal to the 70th percentile of the book-to-market ratio of all NYSE traded stocks are considered as *value* stocks.

In the next step, Fama and French then combine the two sorts by performing a "double sort". This implies, e.g., that they sort the *small* stocks into *small-growth*, *small-neutral*, and *small-value* portfolios. Overall there are thus the following six portfolios:

			Book-to-market ratio		
			low *growth*	medium *neutral*	high *value*
Size	low	*small*	(1)	(2)	(3)
	high	*big*	(4)	(5)	(6)

Note that the number of firms within each of the six portfolio will be different from the other portfolios.

The next step of Fama and French is then to compute the (simple) return on each of the six portfolios. To calculate these, you need the stock returns of the firms contained in that portfolio and the portfolio weight of each stock. Fama and French compute portfolio returns using portfolio weights which reflect a stock's market cap relative to the market cap of the portfolio. These so-called *value-weighted portfolio returns* are reported in the "6 Portfolios Formed on Size and Book-to-Market (2 x 3)" file that's available on Kenneth French's homepage.

To get a rough idea how value-weighted portfolio returns are calculated, consider the following example. Suppose there are $n = 2$ stocks in a portfolio. The first stock has a return of 0.06 and a market capitalization of 70\$, whereas the second stock yields a return of 0.18 and has a market capitalization of 30\$. Consequently, the total market cap of the portfolio is 70\$ + 30\$ = 100\$. Thus the portfolio weights for the two stocks are $\frac{70\$}{100\$} = \frac{7}{10}$ and $\frac{30\$}{100\$} = \frac{3}{10}$. Therefore, the *value-weighted* portfolio return is $\frac{7}{10} \cdot 0.06 + \frac{3}{10} \cdot 0.18 = 0.0960$.[5]

To focus on the difference between small and big stocks, Fama and French construct the so-called "size portfolio SMB (Small Minus Big)". Therefore, they calculate the *average* (simple) return on the *small* portfolio and on the *big* portfolio:

$$\begin{aligned} \text{small} &= \frac{1}{3}\,(\text{small growth} + \text{small neutral} + \text{small value}) \\ \text{big} &= \frac{1}{3}\,(\text{big growth} + \text{big neutral} + \text{big value})\,. \end{aligned}$$

The Fama and French factor *Small Minus Big (SMB)* (or "size factor") is the result of a trading strategy that involves going long in the *small* portfolio (containing stocks with *low* market capitalization) and going short in *big* portfolio (containing stocks with *high* market capitalization). Put differently, this trading strategy involves buying the small stocks and selling the big stocks.

Fama and French also focus on the difference between stocks with low and high book-to-market ratios. Hence, they compute the *average* (simple) return on the *growth* portfolio and on the *value*

[5] An alternative would be *equal-weighted* portfolio returns. In this case, all n stocks which enter a portfolio have the same weight, $\frac{1}{n}$. Consider again the example for $n = 2$ from above. The equal weighted portfolio return is $\frac{1}{2} \cdot 0.06 + \frac{1}{2} \cdot 0.18 = 0.1200$.

portfolio:

$$\begin{aligned} \text{growth} &= \frac{1}{2}\,(\text{small growth} + \text{big growth}) \\ \text{value} &= \frac{1}{2}\,(\text{small value} + \text{big value})\,. \end{aligned}$$

The Fama and French factor *High Minus Low (HML)* (or "value factor") is the result of a trading strategy that involves going long in the *value* portfolio (containing stocks with *high* book-to-market ratios) and going short in the *growth* portfolio (containing stocks with *low* book-to-market ratios). Put differently, this trading strategy involves buying value stocks and selling growth stocks.

C MATLAB for financial econometrics toolbox

C.1 Installing the MFE toolbox

1. Open this homepage and download the MFEToolbox.zip. You can also find a pdf file with an extensive documentation of the corresponding functions on this homepage.
2. Unzip the file MFEToolbox.zip.[6]
3. A new folder "MFEToolbox" is created. It contains (among others) the MATLAB file `addToPath.m`. Open it in MATLAB. As soon as the editor opens, click the "Run" button.
4. Switch to MATLAB's command window. The remaining installation steps just require you to answer questions posed in the command window by entering yes (`Y`) or no (`N`).
5. `Add MATLAB work-a-like functions? [Y/N(default)]`. Please enter `Y`.
6. `Make path permanent? [Y/N (default)]`. Please enter `Y`.
7. `Add MEX files to path? [Y(default)/N]`. Please enter `Y`. (You will only get that question if you are using a Windows computer.)
8. Now you can use the function `olsnw` in the same way as any of MATLAB's built-in-functions.

C.2 Using the `olsnw` function

The `olsnw` function estimates a linear regression with either White (1980)'s heteroscedasticity consistent standard errors or Newey and West (1987)'s heteroscedasticity and autocorrelation consistent (HAC) standard errors. The function has the following syntax:

```
[b, tstat, ~, ~, ~, Rbar, ~] = olsnw(y,x,c,nwlags)
```

- Input variables:
  - `y`: dependent data
  - `x`: independent data
  - `c`: indicates whether a constant should included (`1`: include constant, `0`: do not include constant)

[6] You can use, e.g., the program 7-Zip to do that.

  - `nwlags`: indicates the number of lags to correct for autocorrelation (if `0`: White (1980)'s heteroscedasticity consistent standard errors)

- Output variables:
  - `b`: vector of parameter estimates
  - `tstat`: vector of t-statistics
  - `Rbar`: adjusted R^2