Casper Krusberg 187410 & Robin Perälä 175910

Assignment 1

Empirical Methods in Finance 17011

1. A student named Abu is writing a term paper on a statistics course in Hanken. He realizes that he doesn't know how to code a particular data set chart, and asks his friend on the course, a student called Dabu, if he could send a copy of his graph, displaying the graph solution as advice. Abu decides to copy-paste the same graph into his term paper, without using his friend Dabu as a reference. He assumes the great professor Dr. Jan Antell, won't notice as it isn't possible to google image search the same graph, and the amount of term paper graphs Antell will receive will look very similar anyways. Plagiarism has taken place.

2. **Background information:** We have the annual returns of assets A, B and C, from a six-year time period. From those returns, we calculate the variances, and the covariances with respect to each other. We present them in matrix-form, called the covariance matrix (also called variance-covariance matrix). Our estimation for the expected returns is based on the annual geometric mean for the time period. We summarize their expected returns inside an expected returns vector. The expected returns and the covariance matrix are structured in the following fashion:

$$\bar{\mathbf{R}} = \begin{bmatrix} \bar{R}_1 \\ \bar{R}_2 \\ \vdots \\ \bar{R}_n \end{bmatrix} \qquad \mathbf{V} = \begin{bmatrix} S_1^2 & S_{12} & S_{13} \\ S_{21} & S_2^2 & S_{23} \\ S_{31} & S_{32} & S_3^2 \end{bmatrix},$$

(Antell, J. Microsoft PowerPoint - emf_2021_lectures.pptx. p.78)

where $\bar{\mathbf{R}}$ is the vector of expected returns and $\mathbf{V}$ is the covariance matrix. $\bar{R}_n$ is the expected return of asset n, S_n is the variance of asset n and S_{nm} is the covariance of assets n and m.

Our matrices are the following:

$$expected\ returns\ = \begin{bmatrix} 1.84 \\ 0.95 \\ 4.38 \end{bmatrix}$$

(Numbers are in annual percentages)

$$covariance\ matrix = \begin{bmatrix} 38.22 & 5.74 & -1.62 \\ 5.74 & 119.67 & -36.90 \\ -1.62 & -36.90 & 21.43 \end{bmatrix}$$

The risk-free rate is set to be = 1.33%

A. **A correct covariance matrix is always symmetric and positive semi-definite.** The symmetricity can be confirmed by looking at the elements (values) inside the covariance matrix (i.e., the element for cell a_{21} is the same as a_{12}, the element for cell a_{32} is the same as a_{23}, etc.). In R we checked that the covariance matrix was symmetric using the function "isSymmetric()".

The positive semi-definiteness of a symmetric matrix can be checked through the matrix's eigenvalues. If all eigenvalues calculated of the matrix are non-negative, the matrix is positive semi-definite. In R this is done through the function "eigen()". The eigenvalues of our covariance matrix are all positive (checked with function "all(eigen()>=0)") indicating that the matrix indeed is positive semi-definite. This is also double-checked in our code through the function "is.positive.semi.definite()" available from package "Matrixcalc" (Novomestky 2021). Not only is our matrix semi-definite, but it is also positive definite. If all eigenvalues are bigger than 0, the matrix is positive definite.

Why semi-definite: If we have a covariance matrix "V" of dimension *m* x *m* and a matrix "a" of dimension *m* x 1. Then the variance of a portfolio is defined as a^TVa, where "a" is the vector of the weights (we could also use the notation w^TVw). **Semi-definiteness of a matrix means that a^TVa is non-negative for any a. This must hold for any real numbers, otherwise V is not a covariance matrix and a^TVa is not the variance.** This makes sure that the diagonal elements are all positive. It also makes sure that the covariances are valid and "make sense", meaning that e.g. if the correlation between assets A and B is 1 and A and C is 1, then the correlation between B and C must be 1 as well. Or if the correlation between A and B is 1 but the correlation between A and C is -1, then the correlation between B and C must be -1. There are more detailed mathematical explanations for this.

B. The weights of the minimum variance portfolio we calculate using the following formula:

$$\underbrace{\mathbf{w}_{\min}}_{n\times 1} = \frac{\mathbf{V}^{-1}\mathbf{1}_n}{\mathbf{1}_n^T\mathbf{V}^{-1}\mathbf{1}_n}$$

(Antell, J. Microsoft PowerPoint - emf_2021_lectures.pptx. p.38)

Where V is the covariance matrix and 1_n is a vector of ones, of size n. The superscript "T" is a notation for the transpose and the superscript "-1" is a notation for the inverse. Using the formula in R, we receive the following weights:

$$\textbf{minimum variance portfolio weights} = \begin{bmatrix} 0.122 \\ 0.234 \\ 0.644 \end{bmatrix}$$

Where 0.122 is for asset A, 0.234 for asset B, and 0.644 for asset C

Sum equals 1

We created a minVarWeights() function that calculates the minimum variance portfolio weights, using the earlier formula. In addition to this, we also checked if it's possible to obtain an even smaller variance when shorting is allowed. This was done using the function minvar() of the package "NMOF". As parameters we added wmin = -Inf & wmax=Inf, which indicates that shorting is allowed. The minimum variance portfolio was the same with and without shorting, i.e. none of the assets were given a negative weight.

C. The Sharpe ratio for the portfolio is calculated to be 0.87 through the following formula:

$$\text{Sharpe}_p = \frac{\mathrm{E}(R_p) - R_f}{\sigma_p}$$

Where $E(R_p)$ is the expected return of the portfolio, R_f is the risk-free rate and σ_p is the standard deviation of portfolio p.

3. The determinants of the historical returns for the smoke industry:

We try to estimate the determinants that have impacted the returns for a US smoke industry portfolio, using the Fama-French five-factor model.

We chose to examine the smoke industry's returns, since when comparing to other industries, the smoke industry has had historically high returns. This was interesting to us, considering the popularity of ESG responsible investments nowadays and the smoke industry is not seen as part of them. The reason for choosing the Fama-French 5 factor model for our regression was because the data extraction was straightforward, the model is a widely accepted and often used model, and the model was relevant for our purpose. The Fama-French 5 factor model is an extended version of the Capital Asset Pricing Model.

A. The data has been extracted from the Fama-French data library
available:https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html
We used data from the time period July 1963- August 2021 for our estimation.
Data type: Panel data
Fama/French 5 Factors (monthly arithmetic returns)
49 Industry Portfolios (monthly arithmetic returns)
More information about the data can be found when clicking the link, and then clicking on any portfolio's "details". The returns were collected in American dollars but are stated as percentages.

B. The main research question is:

- How well does the Fama-French five-factor model explain the smoke industry's returns during the time period July 1963- August 2021?

Sub questions:

- How much of the variation in the smoke industry's returns, can be explained by the Fama-French five-factor model's R^2?
- How well does each explanatory variable explain the returns? Are they significant and on what level?

The variables of interest are $\mathbf{R_m}$ (market risk premium), **SMB** (the company size effect on returns), **HML** (the company book-to-market value effect on returns), **RMW** (Robust-Minus-Weak portfolio effect on returns) as well as **CMA** (Conservative Minus Aggressive portfolio effect on returns. Our dependent variable is the ***industry return.***

As we are studying how well the whole Fama-French five-factor model explains the returns, **all 5 factors are variables of interest, and thus we have no control variables.** An example of using control variables would be if we would have begun by using only the Fama-French three-factor model, and then added RMW and CMA as control variables to check if the model would have improved.

C. We used the following method:

We extracted the data from the industry returns for the smoke industry for the time period July 1963- August 2021, and turned it into a vector in R. We then made vectors in R out of each "explanatory variable column" from the Fama-French five-factor model excel file. We continued by checking that the data sets are equal in terms of time periods, and then we constructed the linear regression model in R

$$R_{it} - R_{ft} = \alpha + \beta_{R_m} R_m + \beta_{SMB} SMB + \beta_{HML} HML + \beta_{RMW} RMW + \beta_{CMA} CMA + \epsilon_{it}$$

Where:
$R_{it} - R_{Ft}$ is the excess return in month t for the smoke industry portfolio
R_{Ft} is the risk-free rate
R_m is the market risk premium
SMB is the return spread of small minus large stocks (i.e., the size effect)
HML is the return spread of cheap minus expensive stocks (i.e., the wealth effect)
RMW is the return spread of the firms most profitable firms minus the least profitable
CMA is the return spread of firms that invest conservatively minus aggressively

For checking if the variables of interest are significantly explaining the smoke industry returns, we create the following hypotheses. Null hypothesis: the Fama-French factors have no relationship with the smoke industry returns. Alternative: the factors have a relationship with the smoke industry returns. The statistical hypotheses are the following:

H_o: $\beta_{R_m} = 0$	H_1: $\beta_{R_m} \neq 0$
H_o: $\beta_{SMB} = 0$	H_1: $\beta_{SMB} \neq 0$
H_o: $\beta_{HML} = 0$	H_1: $\beta_{HML} \neq 0$
H_o: $\beta_{RMW} = 0$	H_1: $\beta_{RMW} \neq 0$
H_o: $\beta_{CMA} = 0$	H_1: $\beta_{CMA} \neq 0$

We use the t-test to check these. Because we have a large sample size (697 data points), the t-distribution has converged close to the normal distribution.

For checking if the model is significantly explaining the smoke industry returns, we create the following hypotheses. Null hypothesis: the Fama-French five-factor model has no relationship with the smoke industry returns. Alternative: the model has a relationship with the smoke industry returns. The statistical hypotheses are the following:

H_0: $\beta_{R_m} = \beta_{SMB} = \beta_{HML} = \beta_{RMW} = \beta_{CMA} = 0$
H_1: $\beta_i \neq 0 \; for \; \exists i,$

Where *i* are the five factors. In words: the null hypothesis is that all coefficients are equal to zero and the alternative is that at least one coefficient is statistically significantly different from zero. We test this using the F-test and F-distribution, which is used for e.g. comparing two models against each other. Our models are the Null and the alternative hypothesis.

We obtain T-test and F-test values in R, and we can check the significance from the p-values.

Lastly, we ran tests on our linear regression model, to test for heteroscedasticity, multicollinearity, autocorrelation, and normality. This was carried out through the tests: Whites-test, Jarque-Bera and Breusch-Godfrid. We also looked the Variance inflation factors (auxiliary regression) and the correlation matrix.

Preliminary relationship between the variables (What we expect):
The all focus on portfolio returns and some of them might have quite high correlations, therefore it is important to check this through a correlation matrix.

- We expect general market return to correlate positively with the smoke industry returns.
- We expect SMB to have a somewhat negative correlation with the smoke industry returns since smoke companies are usually quite large by nature.
- We expect HML to have a close to 0 correlation to the smoke industries returns since smoke companies' stocks aren't very cheap but not very expensive either.
- We expect RMW to have a somewhat positive correlation since the smoke industry seems fairly profitable
- We expect CMA to have a somewhat positive correlation since the smoke industry seems like quite a conservative investing class.

D. Results, descriptive statistics, and diagnostics

We found that our linear regression model had 3 significant variables explaining the smoke industry returns. These were the following:

R_m	- on a 1% significance level	coefficient: 0.853
RMW	- on a 1% significance level	coefficient: 0.669
CMA	- on a 1% significance level	coefficient: 0.769

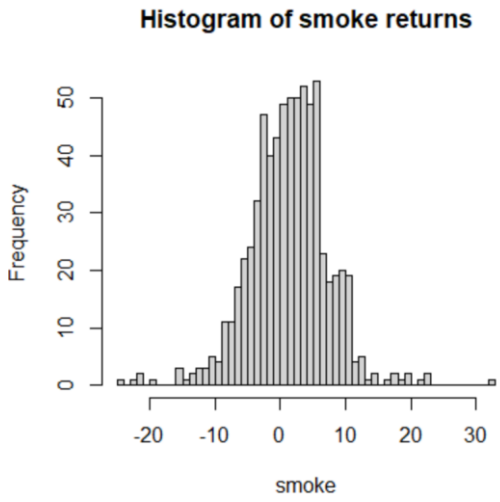
SMB and **HML** were not significant even on a 10% level. coefficients: -0.094 & -0.137

The **R^2** value was **0.322**, and the **adjusted R^2** was **0.317**, indicating that our model did in fact not explain the smoke industry returns that well. **This even though the model in itself <u>was significant</u>, with a p-value of less than 1%.**

The model intercept had a value of 0.144. This wasn't however a significant result.

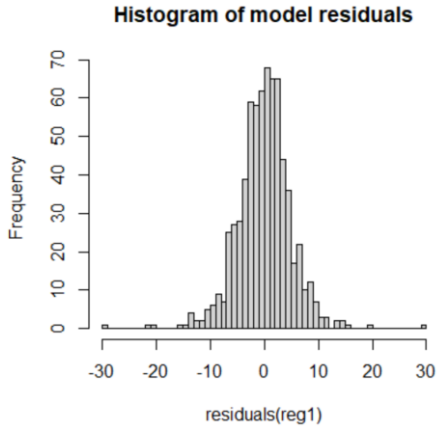
<u>Descriptive statistics</u>

	Smoke	R_m	SMB	HML	RMW	CMA
Minimum	-24.93	-23.24	-15.39	-14.02	-18.76	-6.78
Median	1.66	0.93	0.12	0.23	0.24	0.10
Mean	1.33	0.58	0.24	0.26	0.26	0.26
Geometric mean monthly	1.14	0.48	0.19	0.23	0.24	0.24
Geometric mean annual	14.63	5.96	2.31	2.70	2.88	2.96
Maximum	32.47	16.10	18.38	12.48	13.38	9.06



We checked for multicollinearity by creating a correlation matrix and checking individual values. We found that variables HML and CMA have quite a "large" correlation (value: 0.6697). After that we tested for multicollinearity through the VIF test and once again received higher values for HML (1.89) and CMA (2.13) compared to the other variables. This indicates moderate association between the explanatory varibales from a VIF point-of-view. We tried to leave out variable HML in our regression (since it wasn't significant) and see if the model will explain the scenario better or worse. We found that the new $\mathbf{R^2}$ received a value of **0.319**, and adjusted $\mathbf{R^2}$ **0.315.** The model in itself remained significant, but the model DID NOT get better according to the adjusted R^2.

We test for normality in the error terms of our model. This is conducted in R through the Jarque-Bera normality test. If the test result is far from zero, it signals the residuals are not normally distributed. The test has a value of 565.7 indicating that the residuals are not normally distributed. The test was significant on a 1% level. After printing a histogram of our model's residuals, we in fact see that there are some outliers. Our sample is however fairly large (697 observations), which should make our statistics more normal according to the Central limit theorem.



We try to improve the properties of our data by making a logarithmic transformation of the returns (in hope this would make the distribution more normal) but find that our model does not improve by doing this. The residuals are as a matter of fact now even less normally distributed. The non-normally distributed residuals can have been affected by our data (returns) also not being completely normally distributed, this can be seen from the histogram below. One way to correct for this could have been by removing data outliers through winzorisation. However, in this model we decided to keep the data outliers as is, as they were not very extreme.

We test whether the variance of the residuals is inequal over the range of measured values (known as heteroskedasticity). We do this using White's test and get a p-value of approximately 0.45. Knowing this, we reject the null hypothesis in White's test. It seems like there is heteroskedasticity in our model. We also use the simpler Breusch-Pagan test (to confirm our heteroskedasticity suspicions) and get the same result. The only different between White's test and the Breusch-Pagan is that the Breusch-Pagan auxiliary regression doesn't include cross-terms or the original squared variables. We used White's test because our sample was quite large. The Breusch-Pagan test gives a value of 4.71, and a p-value of 0.4522. We reject the Breusch-Pagan null hypothesis.

We try to find out in which variable the problem lies. We summarize White's test in R and see that the interactions between R_m: RMW and RMW: CMA are significant on all general levels. HML: RMW is significant on a 5% significance level. We assume that it's the variable RMW that is affecting the heteroskedasticity. We try removing RMW from our original model but find that the (adjusted) R2 doesn't increase (model is still significant). The RMW variable was significant on all general levels in the original model, and it improves the model slightly, so we decide not to remove it. We instead try to correct for heteroskedasticity by using robust standard errors (Heteroskedasticity-consistent standard errors are used to allow the fitting of a model that does contain heteroskedastic residuals). For the White robust standard errors: There are 4 heteroskedasticity-consistent (HC) standard error formulas; HC1, HC2, HC3 and HC4. We decide to use HC3 as our sample is large and we expect heteroskedasticity. After having corrected for heteroskedasticity, we see that R_m, RMW, CMA (and our model overall) are still significant on all general significance levels.

We expect autocorrelation in the error terms, because monthly returns usually include some momentum behaviour. We decide to test for this. This is done in R through the Breusch-Godfrey test. The null hypothesis is that there is no serial correlation of any order up to *p*. We set *p* (lags) to 12 since we are dealing with monthly returns. We receive a test value of 13.452 and a p-value of 0.3371. We don't reject the null hypothesis, meaning that there seems to be little autocorrelation in the error terms. If there would have been problems, we could have decreased the adverse effects of autocorrelation by using the robust standard errors (Andrew's method). We still check what would happen if we would correct for autocorrelation. The results show that R_m, RMW and CMA still are significant on all general significance levels. We test once more for the significance of the whole model and find that it still indeed is significant on all general levels.

Summary of models used, including the original model and tested corrections:

	Five-factor	Without HML	Logarithmic
AIC	4236.96	4237.29	4238.14
BIC	4268.79	4264.57	4269.97
Adjusted R^2	0.3167	0.3154	0.3155

The five-factor model was the best in terms of all three statistics

E. We found indications that the independent variables SMB, CMA & R_m are significant in explaining the returns for the smoke industry for the time period in question.

The independent variable HML did not explain the smoke industry's returns in a significant way.

The linear regression model in itself was significant on a 1% significance level.

F. The linear regression model test conducted in this assignment could be used as an example in the beginning of the EMF course, as repetition for both those who attended the econometrics course and the basic course in statistics.

The model in itself is easy to understand and can be modified in different ways. What is important to remember, is to test the usability of the model each time it is modified, as it after modifying might have problems with homoscedasticity, multicollinearity, autocorrelation, and normality.

We also included tests for heteroscedasticity, multicollinearity, autocorrelation, and normality for our model. Even though these tests aren't explained in great detail, we have explained why we have used them, and it can therefore work as a small reminder in the beginning of the course for students who don't remember or aren't familiar with these tests' purposes.

Extra example for exercise 1 (optional reading):

Pekka: The serial cheater

Pekka is a serial cheater. Everywhere he goes, he cheats. During non-pandemic times he has all his tools in this tool belt available. Today there is an exam at Hanken. Pekka has written answers on his arms, legs, and belt. He can read them at will. He also has a small high-tech smart gadget with him, where he can google anything. It is integrated into the lens of his spectacles with a small partly transparent screen. If the spectacles fail, he has a reserve gadget built into his shirt. Both gadgets can be used with touch or voice control. The newest technology has also come to town, so that he can control them, by just using the movement of his eyes. In a coming update it can simply be controlled through the electromagnetic waves in his brain. The gadget will be connected to the nervous system. There was a metal detector at the front door, but fortunately this gadget is anti-detector. Pekka always stays one step ahead the administrators in his cheating career. He always gets the latest and greatest cheating tool. He has any and every tool you can imagine.

But one day, the course provides incentives which stop cheating, and the exam is made in a way where google is of no use. The exam requires some innovative thinking and deep understanding. It is not multiple-choice questions as always. Instead: essay style question! Pekka cannot stand essay style questions. He is hit by a crippling depression, and he ends his cheating career. He crawls back into the cave he came from. Pekka is defeated.

References (in order of appearance in the r script)

Frederick Novomestky (2021). matrixcalc: Collection of Functions for Matrix Calculations. R package version 1.0-5. https://CRAN.R-project.org/package=matrixcalc

Michael Friendly, John Fox and Phil Chalmers (2021). matlib: Matrix Functions for Teaching and Learning Linear Algebra and Multivariate Statistics. R package version 0.9.5. https://CRAN.R-project.org/package=matlib

Manfred Gilli, Dietmar Maringer and Enrico Schumann. Numerical Methods and Optimization in Finance. 2nd edition. Elsevier/Academic Press, 2019.

Enrico Schumann. Numerical Methods and Optimization in Finance (NMOF). Manual. Package version 2.4-1.

Hadley Wickham and Jennifer Bryan (2019). readxl: Read Excel Files. R package version 1.3.1. https://CRAN.R-project.org/package=readxl

John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Lukasz Komsta and Frederick Novomestky (2015). moments: Moments, cumulants, skewness, kurtosis and related tests. R package version 0.14. https://CRAN.R-project.org/package=moments

Achim Zeileis, Torsten Hothorn (2002). Diagnostic Checking in Regression Relationships. R News 2(3), 7-10. URL https://CRAN.R-project.org/doc/Rnews/

Zeileis A, Köll S, Graham N (2020). “Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R.” _Journal of Statistical Software_, *95*(1), 1-36. doi: 10.18637/jss.v095.i01 (URL: https://doi.org/10.18637/jss.v095.i01).

Zeileis A (2004). “Econometric Computing with HC and HAC Covariance Matrix Estimators.” _Journal of Statistical Software_, *11*(10), 1-17. doi: 10.18637/jss.v011.i10 (URL: https://doi.org/10.18637/jss.v011.i10).

Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.1. https://CRAN.R-project.org/package=stargazer