Project 2: Trading ETFs II

Formalities, structure and expectations – second mandatory project

In this project, we'll continue to analyse ETF return data. It is generally believed that large returns are associated with large risk. In the following, we will formulate and select a suitable multiple linear regression model for the "average" weekly return of the ETFs, which has selected risk measures as explanatory variables. The assignment must be solved using the statistical software R. Some code suggestions are provided but, in addition, it's a good idea to take a look at the R code from project 1, as well as chapter 5 and 6 of the book.

The results of your analysis must be documented in a report with tables, figures, appropriate mathematical notation, and explanatory text. Relevant figures and tables must be included within the text, not in an appendix. Present the results of your analysis as you would when explaining them to one of your peers. Divide the report into subsections, one for each of the questions to be answered.

The report must be handed in as a pdf file. R code should not be included in the report itself, but must be handed in as an appendix (a .R file). The report and appendix must be handed in under Opgaver/Assignments on CampusNet at:
`Assignments > Active Assignments > Obligatorisk opgave nr. 2:`
`Handel med ETF II > Answer > Answer Assignment`

The report should not exceed 6 pages (excluding figures, tables, and the appendix). A page contains 2400 characters.

It's important that you describe and explain the R output in words – figures and tables cannot stand alone.

When you're asked to state a formula, insert numbers, and then perform certain computations, it's important to show that you've done this by including your intermediate results. (In these cases, it's not enough to report results obtained directly from R). Furthermore, remember that when performing a hypothesis test, you must go through the following steps: State the hypothesis and significance level (α), compute the test statistic and state it's distribution, compute the p-value, and summarize your findings.

Figures and tables are not included in the assessment of the length of the report. However, it's not in itself an advantage to include many figures, if they aren't relevant!

You may work in groups, but the report must be written individually. Questions may be addressed to the teaching assistants, see the guidelines on the *Projects* page of the course website.

Data

Read the dataset `finans2_data.csv` into R. The following code may be used:

```
# Read the dataset 'finans2_data.csv' into R
D <- read.table("finans2_data.csv", header = TRUE, sep = ";")
```

The dataset for project 1 contained observations of the weekly return from 95 ETFs during the period 2006-05-05 to 2015-05-08 (altogether 454 weeks). The dataset for project 2 is based on data from the same ETFs, and from the same time period. This new dataset consists of observations of 4 variables: `ETF`, `Geo.mean`, `Volatility`, and `maxTuW`. The variable `ETF` is an id variable, which specifies the name of the ETF in question. The three remaining variables are described in further detail in the following.

The geometric average rate of return

Let X_t denote the price of an ETF at the end of week t. Then, X_{t-1} denotes the price at the end of week $t-1$. Define

$$a_t = 1 + r_t = \frac{X_t}{X_{t-1}},$$

where r_t is the return (the relative return) for the t'th week. The *geometric average rate of return* over the 454 weeks covered by the data is defined as

$$r_{\text{week}} = \sqrt[454]{a_1 \cdot a_2 \cdot \ldots \cdot a_{454}} - 1\,.$$

The variable `Geo.mean` contains observations of the geometric average rate of return for the ETFs in the dataset.

Risk measures

Many different risk measures exist for financial instruments. For example, a very simple measure is the standard deviation of the price of the ETF (not computed for our data). The dataset for this assignment contains the following risk measures:

- *Weekly volatility* (the variable `Volatility`). The standard deviation of the ratio between the price of an ETF at the beginning and end of a week (that is, a_t).

$$v = 100 \cdot \sqrt{\frac{1}{454} \cdot \sum_{t=1}^{454} (a_t - \bar{a})^2},$$

  where $\bar{a} = (a_1 + a_2 + \ldots + a_{454})/454$.

- *"Maximum Time under Water" (maxTuW)* (the variable `maxTuW`, sometimes referred to as "*Maximum Drawdown Duration*"). Indicates the maximum number of weeks between two peak prices.

Statistical analysis

a) Present a short descriptive analysis and summary of the data for the variables `Geo.mean`, `Volatility`, and `maxTuW`. Include scatter plots of the geometric average rate of return against the two other variables, as well as histograms and box plots of all three variables. Present a table containing summary statistics, which includes the number of observations, and the sample mean, standard deviation, median, and 0.25 and 0.75 quantiles for each variable.

The dataset contains observations from altogether 95 ETFs. However, in the following, the statistical model should only be fitted to the data from 91 of these ETFs. Later on, we'll use the data from the four remaining ETFs to evaluate the prediction capabilities of the final model. The four ETFs which are to be excluded from the model, are the same ETFs that were in focus in project 1: AGG, VAW, IWN, and SPY. For example, the following code may be used to split the dataset into two parts, one for estimating the model (`D_model`), and the other for validating prediction accuracy (`D_test`):

```
# Subset containing only AGG, VAW, IWN and SPY (for validation)
D_test <- subset(D, ETF %in% c("AGG","VAW","IWN","SPY"))

# Subset containing only the 91 remaining ETFs (for model estimation)
D_model <- subset(D, !(ETF %in% c("AGG","VAW","IWN","SPY")))
```

b) Formulate a multiple linear regression model with the geometric average rate of return as the dependent/outcome variable (Y_i), and volatility and maxTuW as the independent/explanatory variables ($x_{1,i}$ and $x_{2,i}$, respectively). Remember to state the model assumptions. (See equation (6-1) and example 6.1).

c) Estimate the parameters of the model. These consist of the regression coefficients, which we denote by β_0, β_1, β_2, and the variance of the residuals, σ^2. You may use the following R code:

```
# Estimate multiple linear regression model
fit <- lm(Geo.mean ~ Volatility + maxTuW, data = D_model)

# Show parameter estimates etc.
summary(fit)
```

Give an interpretation of the estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$, explaining what they tell us about the relation between the geometric average rate of return and the model's explanatory variables. (See remark 6.14). Furthermore, present the estimated standard deviations of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$, the degrees of freedom used for the estimated residual variance $\hat{\sigma}^2$, and the explained variation, R^2.

d) Perform model validation with the purpose of assessing whether the model assumptions hold. Use the plots, which can be made using the R code below, as a starting point for your assessment. (See section 6.4 on residual analysis).

```
# Plots for model validation

# Observations against fitted values
plot(fit$fitted.values, D_model$Geo.mean, xlab = "Fitted values",
       ylab = "Geom. average rate of return")

# Residuals against each of the explanatory variables
plot(D_model$EXPLANATORY_VARIABLE, fit$residuals,
        xlab = "INSERT TEXT", ylab = "Residuals")

# Residuals against fitted values
plot(fit$fitted.values, fit$residuals, xlab = "Fitted values",
```

```
     ylab = "Residuals")

# Normal QQ-plot of the residuals
qqnorm(fit$residuals, ylab = "Residuals", xlab = "Z-scores",
       main = "")
qqline(fit$residuals)
```

e) State the formula for a 95% confidence interval for the volatility coefficient, here denoted by β_1. (See method 6.5). Insert numbers into the formula, and compute the confidence interval. Use the R code below to check your result, and to determine confidence intervals for the two other regression coefficients.

```
# Confidence intervals for the model coefficients
confint(fit, level = 0.95)
```

f) It is of interest whether β_1 might be -0.06. Formulate the corresponding hypothesis. Use the significance level $\alpha = 0.05$. State the formula for the relevant test statistic (see method 6.4), insert numbers, and compute the test statistic. State the distribution of the test statistic (including the degrees of freedom), compute the p-value, and write a conclusion.

g) Use backward selection to investigate whether the model can be reduced. (See example 6.13). Remember to estimate the model again, if it can be reduced. State the final model, including estimates of its parameters.

h) Use your final model from the previous question as a starting point. Determine predictions and 95% prediction intervals for the geometric average rate of return, for each of the four ETFs in the validation set (`D_test`). See Example 6.8, Method 6.9 and the R code below. Compare the predictions to the observed geometric average rates of return for the four ETFs in the validation set and make an assessment of the prediction capabilities of the final model.

```r
# Predictions and 95% prediction intervals
pred <- predict(FINAL_MODEL, newdata = D_test,
                interval = "prediction", level = 0.95)

# Observed values and predictions
cbind(id = D_test$ETF, Geo.mean = D_test$Geo.mean, pred)
```

Hence, don't write the formulas in the report, but instead refer to that the R function `predict` was used for the calculations. The formulas requires a matrix formulation, which are out of the curriculum (to derive the formulas use Equations (6-48) and (6-49) together with the derivations leading to Equations (5-57) and (5-58)).