Projekt 1: Handel med ETF

Formaliteter, opgavestruktur og forventninger til 1. obligatoriske opgave

Opgaven består af to dele. I første del skal der laves en deskriptiv analyse af data. Anden del handler primært om konfidensintervaller og hypotesetests.

Der er lagt op til, at man skal arbejde med opgaven i små "lette" trin. Opgaven skal i praksis løses ved hjælp af programmet `R`. Der er udarbejdet `R`-kode, som gør det nemt at komme i gang med projektet. Koden er dog ikke fuldstændig, og I opfordres til at udforske R samtidig med at I laver projektet. F.eks. kan I arbejde med at lave "pæne" titler til graferne eller benytte R's indbyggede funktioner til beregning af konfidensintervaller og test af hypoteser.

Besvarelsen skal dokumentere den gennemførte analyse ved tabeller, grafer, matematisk notation, samt tekst der beskriver analysens resultater. Relevante grafer og tabeller skal indgå i sammenhæng med teksten - ikke som bilag. Præsenter resultaterne fra jeres analyser på samme måde, som I ville videreformidle dem til andre fagfæller.

Inddel besvarelsen i et underafsnit for hver af de stillede spørgsmål.

Besvarelsen med bilag skal afleveres som pdf-fil. R-kode bør ikke indgå i besvarelsen, men vedlægges som bilag (i form af en .R-fil). Besvarelsen samt bilag afleveres under Opgaver/Assigments på CampusNet ved: `Opgaver > Aktive opgaver > Obligatorisk opgave nr. 1: Handel med ETF > Besvar > Besvar opgave`

En samlet besvarelse bør ikke overstige 6 sider (ekskl. plots, tabeller og bilag). En normal side udgør 2400 anslag.

Grafer og tabeller kan IKKE stå alene - det er altså vigtigt, at I beskriver og fortolker outputtet fra `R` med ord.

Grafer og tabeller indgår ikke i opgørelsen af besvarelsens længde. Det er dog IKKE i sig selv en fordel at medtage mange plots, hvis de ikke er relevante!

I må gerne arbejde sammen i grupper, men besvarelsen af opgaven skal skrives individuelt. Spørgsmål omkring projektet kan rettes til hjælpelæren, se retningslinjerne på siden `Projects` på kursets hjemmeside.

Problemstilling

I dette projekt ser vi på det ugentlige afkast for et udvalg af ETF'er. En ETF ("Exchange-Traded Fund") kan beskrives som en struktureret, børsnoteret pulje af aktier. ETF'erne kan altså købes og sælges på samme måde som almindelige aktier på en fondsbørs. En ETF er en kollektiv investeringsfond, der ligner en investeringsfond eller investeringsforening. ETF'er besidder således en kombination af fordelene ved kollektive fonde og aktier. Køber man f.eks. en simpel ETF der dækker SP100-indekset i USA svarer det til at eje en del af alle 100 aktier i indekset. Således slipper man for at købe 100 enkeltpapirer, man kan nøjes med at købe et.

Der findes mange forskellige ETF'er – faktisk er ETF-markedet under eksplosiv udvikling. Der er også forskellige strategier, hvorunder ETF'erne administreres. En ETF med passiv strategi tilstræber at følge afkastet i det underliggende indeks så tæt som muligt. Her er formålet altså at give investorerne det samme afkast som det underliggende marked. Denne type ETF kaldes en indeksfond. Et eksempel kunne være EURO STOXX-indekset, som indeholder de førende aktier i Eurozonen. Hvis EURO STOXX50-indekset f.eks. stiger med 10% i løbet af et år skal en ETF, der følger dette indeks, tilstræbe at give investorerne det samme afkast, minus gebyrer, som hos ETF'er benævnes *det totale omkostningsforhold* (TER). For at kunne levere det samme afkast som markedsindekset indeholder ETF'en de samme værdipapirer eller en repræsentativ del af de værdipapirer, der er i indekset.

Fordelen ved ETF'er i forhold til f.eks. investeringsforeninger er fleksibilitet, omkostningseffektivitet og høj likviditet. ETF'erne er altså billige sammenlignet med andre investeringsprodukter. Som med alle investeringer, er der dog en risiko forbundet med køb og salg af ETF'er. Værdien af investeringen kan falde, og det investerede beløb kan tabes. [1]

[1] Ovenstående afsnit er skrevet på baggrund af information fra: `http://www.ishares.com` og `https:`

Indlæsning af data

Lav en mappe til projektet på din computer. Download materialet til projektet fra CampusNet og udpak ("unzip") det til den mappe, du lige har lavet.

Åbn derefter data-filen `finans1_data.csv` (f.eks. i RStudio, File → Open File) for at se filens indhold. Bemærk at første linje indeholder variabelnavne (kaldes en *header*), og at de efterfølgende linjer indeholder de egentlige observationer. Observationerne af de enkelte variable er adskilt af et ';' (deraf `.csv`: "comma separated values", her dog semikolon).

Observationerne i datasættet består af ugentlige afkast (forholdet mellem slut- og startkurs for pågældende uge minus 1) for 95 ETF'er. I data-filen er der 96 søjler. Første søjle indeholder datoer, mens hver af de resterende 95 søjler indeholder ugentlige afkast for én ETF. Søjlenavnet angiver navnet på den pågældende ETF.

Åbn filen `finans1_dansk.R`, som indeholder R-kode der kan bruges til analysen. Først skal "working directory" sættes til den mappe på computeren, hvor filerne til projektet er gemt:

```
## I RStudio kan man nemt sætte working directory med menuen
## "Session -> Set Working Directory -> To Source File Location"
## Bemærk: i R bruges kun "/" til separering i stier
## (altså ingen backslash).
setwd("Erstat her med stien til den mappe, hvor projektfilerne er gemt.")
```

Nu kan datasættet indlæses i R med følgende kode:

```
## Indlæs data fra finans1_data.csv
D <- read.table("finans1_data.csv", header=TRUE, sep=";", as.is=TRUE)

## Udvælg tids-variablen samt ETF'erne AGG, VAW, IWN og SPY
D <- D[ ,c("t","AGG","VAW","IWN","SPY")]
```

Bemærk, at der i R-koden bruges engelsk. Det er generelt ikke en god ide at bruge æøå osv. ved programmering. `D` bliver en "data.frame" (en slags tabel), som indeholder den indlæste data (se R-introen i kapitel 1.5 i bogen). Der udvælges 4 forskellige ETF'er, *AGG*, *VAW*, *IWN*, og *SPY*, som er de eneste der skal benyttes i den videre analyse. En beskrivelse af de udvalgte ETF'er fremgår af tabellen i bilag 1 på side 10. Filen `ETF_dokumentation.xls` giver en beskrivelse af alle ETF'erne. Tabellen i bilag 1 er hentet fra denne fil.

`//falconinvest.dk/hvad-er-en-etf/`.

Beskrivende analyse (descriptive analysis)

Første del af projektet går ud på at lave en beskrivende analyse af data. I en rapport er det vigtigt at præsentere og beskrive data for læseren. Dette kan f.eks. gøres ved hjælp af opsummerende størrelser/nøgletal ("summary statistics") og passende figurer.

En simpel opsummering af det indlæste datasæt fås ved at køre følgende kode:

```
## Dimensionen af D (antallet af rækker og søjler)
dim(D)
## Søjle-/variabelnavne
names(D)
## De første rækker/observationer
head(D)
## De sidste rækker/observationer
tail(D)
## Udvalgte opsummerende størrelser
summary(D)
## En anden type opsummering af datasættet
str(D)
```

a) Lav en kort beskrivelse af datamaterialet: Hvilke variable indgår i datasættet? Er der tale om *kvantitative* og/eller *kategoriserede* variable (eller dato-variable)? (Kategoriserede variable dukker først op i kapitel 8, men det er bare variable, som inddeler observationerne i kategorier - f.eks. tre kategorier: lav, mellem og høj). Hvor mange observationer er der? Hvilken periode dækker observationerne over (hvornår er første hhv. sidste observation foretaget)? Er der manglende værdier for nogen af variablene?

Et "density histogram" der beskriver den empiriske tæthed af de ugentlige afkast for ETF'en AGG (se kapitel 1.6.1) kan laves ved hjælp af følgende kode:

```
## Histogram der beskriver den empiriske tæthed for afkastene for AGG
## (histogram for AGG-afkast normaliseret så arealet er lig 1)
hist(D$AGG, xlab="Afkast (AGG)", prob=TRUE)
```

b) Lav et density histogram for de ugentlige afkast fra ETF'en AGG. Beskriv fordelingen af afkastene ud fra dette histogram. Er den empiriske tæthed symmetrisk eller skæv? Kan afkastene være både positive og negative? Er der stor spredning i observationerne?

Bemærk: I en *skæv* fordeling er sandsynlighedsmassen ikke symmetrisk fordelt omkring medianen. For en venstreskæv fordeling gælder der, at den længste hale ligger til venstre for midten (almindeligvis vil gennemsnittet også ligge til venstre for medianen). Tilsvarende gælder der, at for en højreskæv fordeling ligger den længste hale til højre for midten (almindeligvis med gennemsnit til højre for medianen).

Ved observationer foretaget regelmæssigt over tid omtales data ofte som en *tidsrække*. Data fra hver ETF udgør således en tidsrække. For tidsrækker er det ofte relevant at lave grafer, der viser udviklingen over tid. Her er det først nødvendigt at fortælle R, at variablen `t` skal opfattes som en dato-variabel. Dette gøres med følgende kode:

```
## Konverterer variablen 't' til en dato-variabel i R
D$t <- as.Date(x=D$t, format="%Y-%m-%d")
## Tjekker resultatet
summary(D$t)
```

Plots der viser hver ETF's udvikling over tid kan nu laves med følgende R-kode:

```
## Plot af ugentligt afkast over tid for hver af de 4 ETF'er
y=c(-0.2,0.2)
## Plot af det ugentlige afkast for ETF'en AGG over tid
plot(D$t, D$AGG, type="l", ylim=y, xlab="Tid", ylab="Afkast (AGG)")
## Tilsvarende plots for de tre andre ETF'er
plot(D$t, D$VAW, type="l", ylim=y, xlab="Tid", ylab="Afkast (VAW)")
plot(D$t, D$IWN, type="l", ylim=y, xlab="Tid", ylab="Afkast (IWN)")
plot(D$t, D$SPY, type="l", ylim=y, xlab="Tid", ylab="Afkast (SPY)")
```

c) Lav plots der illustrerer udviklingen i den ugentlige afkast for hver af ETF'erne. Benyt disse til at beskrive udviklingen i ord. Ser det ud til, at niveauet for afkastet ændrer sig over tid? Er der bestemte tidsperioder, hvor afkastet er bemærkelsesværdigt anderledes? Er der overordnede forskelle i afkastet for de forskellige ETF'er?

Følgende R-kode laver et boxplot over de ugentlige afkast opdelt efter ETF:

```
## Boxplot af afkast opdelt efter ETF
boxplot(D$AGG, D$VAW, D$IWN, D$SPY, names=c("AGG", "VAW", "IWN", "SPY"),
        xlab="ETF", ylab="Afkast")
```

d) Lav et boxplot af afkastene opdelt efter ETF. Benyt derefter plottet til at beskrive den observerede fordeling af afkastene for de fire ETF'er. Er fordelingerne symmetriske eller skæve? Ser det umiddelbart ud til, at der er forskelle mellem fordelingerne (hvis ja, hvilke)? Er der ekstreme observationer/outliers?

Man kan også beskrive den empiriske fordeling af ETF'ernes afkast ved hjælp af opsummerende størrelser/nøgletal som i følgende tabel:

ETF	**Antal obs.**	**Stikprøve-gennem-snit**	**Stikprøve-varians**	**Stikprøve standard-afvigelse**	**Nedre kvartil**	**Median**	**Øvre kvartil**
	n	$(\bar{x})$	(s^2)	(s)	(Q_1)	(Q_2)	(Q_3)
AGG							
VAW							
IWN							
SPY							

For at udfylde de tomme celler i tabellen kan man f.eks. benytte R-kode som følgende (se også bemærkning på side 12 for tricks til udregningerne):

```
## Antal observationer af AGG afkast
## (medregner ej eventuelle manglende værdier)
sum(!is.na(D$AGG))
## Stikprøvegennemsnit for AGG afkast
mean(D$AGG, na.rm=TRUE)
## Stikprøvevarians for AGG afkast
var(D$AGG, na.rm=TRUE)
## osv.
##
## Argumentet 'na.rm=TRUE' sørger for at størrelsen
## udregnes selvom der eventuelt er manglende værdier
```

e) Udfyld tabellen ovenfor med de opsummerende størrelser for de fire ETF'er. Beskriv hvilken ekstra information kan udledes fra tabellen sammenlignet med boxplottet?

Statistisk analyse

Andel del af projektet går ud på at lave en simpel statistisk analyse vedrørende afkastet for ETF'erne. Der skal opstilles statistiske modeller for afkastet. Modellernes parametre skal estimeres, og der skal udføres hypotesetests og beregnes konfidensintervaller.

Konfidensintervaller og hypotesetests

Følgende R-kode kan benyttes til at lave et qq-plot med henblik på at vurdere, om den ugentlige afkast for AGG kan antages at være normalfordelt:

```
## qq-plot for AGG afkast
qqnorm(D$AGG)
qqline(D$AGG)
```

f) Opskriv separate statistiske modeller for det ugentlige afkast fra hver af de fire udvalgte ETF'er (se bemærkning 3.2). Estimer parametrene i de fire modeller (middelværdi og standardafvigelse). Foretag modelkontrol af de antagede forudsætninger (se kapitel 3 samt afsnit 3.1.8 i bogen). Idet konfidensintervaller og hypotesetests her involverer fordelingen af gennemsnit kan det være nyttigt også at inddrage den centrale grænseværdisætning (sætning 3.14) i argumentationen.

I praksis vil der opstå situationer, hvor man på baggrund af f.eks. modelkontrollen *ikke* kan tillade sig at antage, at en models forudsætninger er opfyldte. Da vil man ofte overveje, om det kunne hjælpe at foretage en transformation af data (se kapitel 3.1.9 i bogen). Bemærk at efter en transformation ændres fortolkningen af resultaterne på den oprindelige skala. Det er *ikke* meningen, at I skal lave en transformation af data i dette projekt.

g) Angiv formlen for et 95% konfidensinterval (KI) for det ugentlige middelafkast for AGG (se sektion 3.1.2 i bogen). Indsæt tal og beregn intervallet. Beregn tilsvarende konfidensintervaller for de tre andre ETF'er og udfyld tabellen nedenfor.

	Nedre grænse af KI	**Øvre grænse af KI**
AGG		
VAW		
IWN		
SPY		

Sammenlign det beregnede konfidensinterval for AGG med resultatet af følgende R-kode:

```
## Konfidensinterval for middelafkast for AGG
t.test(D$AGG, conf.level=0.95)$conf.int
```

h) Udfør et hypotesetest med henblik på at undersøge, om middelværdien af det ugentlige afkast for AGG afviger signifikant fra det afkast, som man får ved at gemme pengene under hovedpuden (nemlig ingenting). Dette kan gøres ved at teste følgende hypotese:

$$H_0 : \mu_{\text{AGG}} = 0,$$
$$H_1 : \mu_{\text{AGG}} \neq 0.$$

Angiv signifikansniveauet α, formlen for teststørrelsen samt teststørrelsens fordeling (husk antal frihedsgrader). Indsæt tal, og beregn teststørrelsen og p-værdien. Skriv en konklusion med ord. Kommenter på om det var nødvendigt at udføre det statistiske test, eller om samme konklusion kunne opnås ved konfidensintervallet alene.

Sammenlign resultaterne for test af hypotesen med resultaterne af følgende R-kode:

```
## Test af hypotesen mu=0 for AGG afkast
t.test(D$AGG, mu=0)
```

Vi vil nu undersøge, om der er forskel på det ugentlige afkast for de to ETF'er VAW og AGG.

i) Undersøg ved et hypotesetest, om der kan påvises en forskel mellem middelværdien af det ugentlige afkast for ETF'erne VAW og AGG. Opskriv hypotesen og angiv signifikansniveauet α, formlen for teststørrelsen samt teststørrelsens fordeling (husk antal frihedsgrader). Indsæt tal, og beregn teststørrelsen og p-værdien. Skriv en konklusion med ord. Kan der konstateres en signifikant forskel i afkast mellem VAW og AGG? Hvis ja, for hvilken ETF er afkastet størst?

Sammenlign resultaterne for hypotesetestet med resultaterne af følgende R-kode:

```r
## Sammenligning af afkast for VAW og AGG
t.test(D$VAW, D$AGG)
```

j) Kommenter om det var nødvendigt at udføre hypotesetestet i forrige spørgsmål, eller om samme konklusion kunne opnås ud fra konfidensintervallerne alene? (Se bemærkning 3.59 i bogen).

Korrelation

I forbindelse med opbygning af en portefølje af ETF'er er diversificering/risikospredning et væsentligt begreb. Populært sagt handler det om "ikke at lægge alle sine æg i samme kurv". Risiko kan opgøres på flere måder – f.eks. ved standardafvigelsen på det ugentlige afkast.

Når man vil fastlægge en portefølje af forskellige ETF'er er de parvise korrelationer mellem ETF'er et væsentlig værktøj til at afgøre, hvor meget man vil investere i de forskellige ETF'er. Generelt gælder der, at jo lavere korrelationen mellem de ETF'er som kombineres er, jo mere spredes risikoen. En kombination af ETF'er med negativ korrelation kan bidrage til at mindske volatiliteten (risikoen) af den samlede portefølje.

k) Angiv formlen for beregning af korrelationen mellem ETF'erne VAW og IWN. Indsæt tal og beregn korrelationen. Lav desuden et scatterplot der illustrerer sammenhængen mellem afkastene fra de to ETF'er. Vurder om sammenhængen mellem plottet og korrelationen er som forventet.

Sammenlign den beregnede korrelation med resultatet fra følgende R-kode:

```r
## Beregning af korrelation mellem udvalgte ETF'er
cor(D[ ,c("AGG","VAW","IWN","SPY")], use="pairwise.complete.obs")
```

Bilag 1 Beskrivelse af ETF'erne

Tabellen viser en oversigt over og beskrivelse af de 4 udvalgte ETF'er.

Excel-filen `ETF_dokumentation.xls` giver en beskrivelse af alle ETF'erne. Nedenstående tabel er hentet fra denne fil:

ETF	**Beskrivelse**
AGG	iShares Core Total US Bond Market ETF, formerly iShares Lehman Aggregate Bond Fund (the Fund) seeks investment results that correspond generally to the price and yield performance of the total United States investment-grade bond market as defined by the Lehman Brothers U.S. Aggregate Index (the Index). The Index measures the performance of the United States investment-grade bond market, which includes investment-grade United States Treasury bonds, government-related bonds, investment-grade corporate bonds, mortgage pass-through securities, commercial mortgage-backed securities and asset-backed securities that are publicly offered for sale in the United States. The securities in the Index must have at least one year remaining to maturity. In addition, the securities must be denominated in United States dollars, and must be fixed rate, non-convertible and taxable. The Index is market capitalization weighted. The Fund uses a representative sampling strategy to track the Index. The Fund's investment advisor is Barclays Global Fund Advisors (BGFA).
VAW	Vanguard Materials ETF (the Fund), formerly known as Vanguard Materials VIPERs, is an exchange-traded share class of Vanguard Materials Index Fund, which employs a passive management or indexing investment approach designed to track the performance of the Morgan Stanley Capital International (MSCI) US Investable Market Materials Index (the Index). The Index is an index of stocks of large, medium and small United States companies in the materials sector, as classified under the Global Industry Classification Standard (GICS). This GICS sector is made up of companies in a range of commodity-related manufacturing industries. Included within this sector are companies that manufacture chemicals, construction materials, glass, paper, forest products and related packaging products, as well as metals, minerals and mining companies, including producers of steel. The Fund attempts to replicate the Index by investing all, or substantially all, of its assets in the stocks that make up the Index, holding each stock in approximately the same proportion as its weighting in the Index. The Fund also may sample its target index by holding stocks that, in the aggregate, are intended to approximate the Index in terms of key characteristics, such as price/earnings ratio, earnings growth and dividend yield.
IWN	iShares Russell 2000 Value Index Fund (the Fund) seeks investment results that correspond generally to the price and yield performance of the Russell 2000 Value Index (the Index). The Index measures the performance of the small-capitalization value sector of the United States equity market. It is a subset of the Russell 2000 Index. The Index is a capitalization-weighted index and consists of those companies or portion of a company, with lower price-to-book ratios and lower forecasted growth within the Russell 2000 Index. The Index represents approximately 50% of the total market capitalization of the Russell 2000 Index. The Fund invests in a representative sample of securities included in the Index that collectively has an investment profile similar to the Index. iShares Russell 2000 Value Index Fund's investment advisor is Barclays Global Fund Advisors.
SPY	SPDR Trust, Series 1 (the Trust) is a unit investment trust. The Trust is an exchange-traded fund created to provide investors with the opportunity to purchase a security representing a proportionate undivided interest in a portfolio of securities consisting of substantially all of the common stocks, in substantially the same weighting, which comprise the Standard and Poor's 500 Composite Price Index (the SP Index). Each unit of fractional undivided interest in the Trust is referred to as a Standard and Poor's Depositary Receipt (SPDR). The Trust utilizes a full replication approach. With this approach, all 500 securities of the Index are owned by the Trust in their approximate market capitalization weight.

Bemærkning .1 Ekstra tips til R

Dette er en valgfri ekstra bemærkning om R-kodning (ikke nødvendig for at løse opgaven). Der er mange måder hvorpå man kan udtage en delmængde i R.

```
## Ekstra bemærkning om måder at udtage delmænger i R
##
## En logisk (logical) vektor med sandt (TRUE) eller falsk (FALSE) for
## hver række i D - f.eks:
## De uger hvor der er tab (negativ afkast) på AGG
D$AGG < 0
## Vektoren kan bruges til at udvælge alle de negative AGG afkast
D$AGG[D$AGG < 0]
## Alternativt kan man bruge funktionen 'subset'
subset(D, AGG < 0)
## Mere komplekse logiske udtryk kan laves, f.eks.:
## Find alle observationer fra 2009
subset(D, "2009-01-01" < t & t < "2010-01-01")
```

Bemærkning .2 Ekstra tips til R

Endnu en bemærkning med ekstra R-tips for de interesserede. Man kan f.eks. lave tabellen mere effektivt med en for-løkke.

```
## Lav en for-løkke med beregning af et par opsummerende størrelser
## og gem resultatet i en ny data.frame
num <- 2:5
Tbl <- data.frame()
for(i in num){
  Tbl[i-1,"mean"] <- mean(D[ ,i])
  Tbl[i-1,"var"] <- var(D[ ,i])
}
row.names(Tbl) <- names(D)[num]
## Se hvad der er i Tbl
Tbl

## I R er der endnu mere kortfattede måder sådanne udregninger kan
## udføres. For eksempel
apply(D[, num], 2, mean, na.rm=TRUE)
## eller flere ad gangen i et kald
apply(D[, num], 2, function(x){
  c(mean=mean(x, na.rm=TRUE),
    var=var(x, na.rm=TRUE))
})
## Se flere smarte funktioner med: ?apply, ?aggregate og ?lapply
## og for ekstremt effektiv databehandling se f.eks. pakkerne: dplyr,
## tidyr, reshape2 og ggplot2.

## LaTeX tips:
##
## R-pakken "xtable" kan generere LaTeX tabeller og skrive dem direkte
## ind i en fil, som derefter kan inkluderes i et .tex dokument.
##
## R-pakken "knitr" kan anvendes meget elegant til at lave et .tex
## dokument der inkluderer R koden direkte i dokumentet. Dette
## dokument og bogen er lavet med knitr.
```