Projekt 2: Handel med ETF II

Formaliteter, struktur og forventninger – 2. obligatoriske opgave

I dette projekt fortsætter vi med at analysere data for afkastet fra ETF'er. I almindelighed er der en forventning om, at stort afkast er forbundet med stor risiko. Vi vil derfor opstille en passende multipel lineær regressionsmodel for ETF'ernes "gennemsnitlige" ugentlige afkast, der har udvalgte risikomål som forklarende variable. Opgaven skal i praksis løses ved hjælp af den statistiske software R. Rundt omkring i opgaven er der givet forslag til R-koden, men udover det er det en god idé at se på R-koden fra projekt 1 samt f.eks. kapitel 5 og 6 i bogen.

Besvarelsen skal dokumentere den gennemførte analyse ved tabeller, grafer, passende matematisk notation og 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 hvert af de stillede spørgsmål.

Besvarelsen 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/Assignments på CampusNet ved:
`Opgaver > Aktive opgaver > Obligatorisk opgave nr. 2:`
`Handel med ETF II > Besvar > Besvar opgave`

En samlet besvarelse bør ikke overstige 6 sider (ekskl. plots, tabeller og bilag). En 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.

Når I bliver bedt om at angive en formel, indsætte tal og derefter foretage en beregning er det vigtigt, at I viser I har gjort dette ved at inkludere nogle mellemregninger. (Disse steder er det ikke nok at anføre resultater aflæst i R). Husk også at et hypotesetest består af følgende elementer: Angivelse af hypotese og signifikansniveau (α), teststørrelse inkl. dennes fordeling og p-værdi, samt en konklusion 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.

Data

Indlæs datasættet `finans2_data.csv`. Følgende R-kode kan benyttes:

```
# Indlæs 'finans2_data.csv' filen med data
D <- read.table("finans2_data.csv", header = TRUE, sep = ";")
```

Datasættet til projekt 1 indeholdt observationer af det ugentlige afkast for 95 ETF'er i perioden 2006-05-05 til 2015-05-08 (i alt 454 uger). Datasættet til projekt 2 er baseret på data for de samme ETF'er, og fra den samme tidsperiode. Dette nye datasæt omfatter observationer af 4 variable: `ETF`, `Geo.mean`, `Volatility` og `maxTuW`. Variablen `ETF` er en id-variabel, som angiver navnet på den ETF, som observationen handler om. De tre øvrige variable er nærmere beskrevet i det følgende.

Det gennemsnitlige relative ugentlige afkast

Lad X_t betegne kursen på en ETF ved udgangen af uge t. Således betegner X_{t-1} kursen ved udgangen af uge $t-1$. Definer

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

hvor r_t er afkastet (det relative afkast) for den t'te uge. Det *gennemsnitlige relative ugentlige afkast* (eller det *geometriske gennemsnitsafkast*) over de 454 uger, som data dækker, bestemmes ved

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

Variablen `Geo.mean` indeholder observationer af det gennemsnitlige relative ugentlige afkast for ETF'erne i datasættet.

Risikomål

Der findes mange forskellige risikomål for finansielle instrumenter. Et meget simpelt mål er f.eks. at bruge standardafvigelsen på kursen for en ETF (ikke beregnet for vores data). Datasættet til denne opgave inkluderer følgende risikomål:

- *Ugentlig volatilitet* (variablen `Volatility`). Angiver standardafvigelsen af forholdet mellem kursen på en ETF i begyndelsen og slutningen af en uge (dvs. a_t),

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

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

- *"Maximum Time under Water" (maxTuW)* (variablen `maxTuW`, kaldes i nogle sammenhænge *"Maximum Drawdown Duration"*). Angiver det maksimale antal uger det tager at genopnå en historisk top/peak i ETF'ens kurs.

Statistisk analyse

a) Lav en kort deskriptiv analyse og opsummering af data for variablene `Geo.mean`, `Volatility` og `maxTuW`. Inkluder scatterplots af det gennemsnitlige relative ugentlige afkast mod de to andre variable, samt histogrammer og boxplots af alle tre variable. Der skal også være en tabel med opsummerende størrelser, som for hver variabel inkluderer antal observationer, gennemsnit, standardafvigelse, median samt 25%- og 75%-fraktiler.

Der er observationer fra i alt 95 ETF'er i datasættet. I denne opgave skal den statistiske model dog kun opstilles på baggrund af 91 af disse ETF'er. Observationerne fra de fire andre ETF'er skal vi senere bruge til at vurdere modellens evne til at prædiktere. De fire ETF'er der ekskluderes fra modellen er de samme ETF'er, der var i fokus i projekt 1: AGG, VAW, IWN og SPY. Benyt f.eks. følgende R-kode til at dele datasættet op i et nyt deldatasæt, der benyttes til at estimere modellen (`D_model`), og et der benyttes til validering af modellens prædiktionsevne (`D_test`):

```
# Deldatasæt med AGG, VAW, IWN og SPY (til validering)
D_test <- subset(D, ETF %in% c("AGG","VAW","IWN","SPY"))

# Deldatasæt med kun de øvrige 91 ETF'er (til model)
D_model <- subset(D, !(ETF %in% c("AGG","VAW","IWN","SPY")))
```

b) Opstil en multipel lineær regressionsmodel med det gennemsnitlige relative ugentlige afkast som responsvariabel (Y_i), og med volatilitet og maxTuW som forklarende variable (hhv. $x_{1,i}$ og $x_{2,i}$). Husk at angive forudsætningerne/de statistiske antagelser for modellen. (Se bemærkning 5.6, ligning (6-1) og eksempel 6.1).

c) Estimer modellens parametre, som består af regressionskoefficienterne, her kaldet β_0, β_1, β_2, og residualernes varians, σ^2. Brug evt. følgende R-kode:

```
# Estimer multipel lineær regressionsmodel
fit <- lm(Geo.mean ~ Volatility + maxTuW, data = D_model)

# Vis estimerede parametre mm.
summary(fit)
```

Giv en fortolkning af estimaterne $\hat{\beta}_0$, $\hat{\beta}_1$ og $\hat{\beta}_2$, hvor du forklarer, hvad de siger om relationen mellem det gennemsnitlige relative ugentlige afkast og de to forklarende variable i modellen. (Se bemærkning 6.14). Angiv også de estimerede standardafvigelser for $\hat{\beta}_0$, $\hat{\beta}_1$ og $\hat{\beta}_2$, frihedsgraderne anvendt til estimatet af residualernes varians $\hat{\sigma}^2$, samt modellens forklarede varians, R^2.

d) Foretag modelkontrol for at undersøge, om forudsætningerne for modellen (modellens antagelser) er opfyldte. Benyt de plots, der kan laves ved hjælp af R-koden nedenfor, som udgangspunkt for din vurdering. (Se afsnit 6.4 om residualanalyse).

```
# Plots til modelkontrol

# Observationer mod fittede værdier
plot(fit$fitted.values, D_model$Geo.mean, xlab = "Fittede værdier",
       ylab = "Gnsn. rel. ugentligt afkast")

# Residualer mod hver af de forklarende variable
plot(D_model$FORKLARENDE_VARIABEL, fit$residuals,
        xlab = "INDSÆT TEKST", ylab = "Residualer")

# Residualer mod fittede værdier
plot(fit$fitted.values, fit$residuals, xlab = "Fittede værdier",
```

```
     ylab = "Residualer")

# Normal QQ-plot af residualerne
qqnorm(fit$residuals, ylab = "Residualer", xlab = "Z-scores",
       main = "")
qqline(fit$residuals)
```

e) Angiv formlen for et 95% konfidensinterval for koefficienten for volatilitet, her kaldet β_1. (Se metode 6.5). Indsæt tal i formlen og beregn konfidensintervallet. Benyt derefter nedenstående R-kode til at kontrollere resultatet og til at bestemme konfidensintervaller for de to andre koefficienter i modellen.

```
# Konfidensintervaller for modellens koefficienter
confint(fit, level = 0.95)
```

f) Man er interesseret i, om β_1 kunne have værdien -0.06. Opstil den tilsvarende hypotese. Anvend signifikansniveauet $\alpha = 0.05$. Angiv formlen for den relevante teststørrelse (se metode 6.4), indsæt tal og beregn teststørrelsen. Angiv fordelingen af teststørrelsen (inkl. frihedsgrader), beregn p-værdien og konkluder.

g) Undersøg ved *backward selection* om modellen kan reduceres. (Se eksempel 6.13). Husk at reestimere modellen undervejs, hvis der kan foretages reduktion af modellen. Angiv slutmodellen og estimater for dens parametre.

h) Tag udgangspunkt i din slutmodel fra forrige spørgsmål. Bestem prædikterede værdier og 95% prædiktionsintervaller for det gennemsnitlige relative ugentlige afkast for hver af de fire ETF'er i valideringsdatasættet (`D_test`). Se eksempel 6.8, metode 6.9 og R-koden nedenfor. Sammenlign prædiktionerne med de observerede gennemsnitlige relative ugentlige afkast for disse fire observationer og lav en vurdering af modellens evne til at prædiktere.

```r
# Prædiktioner og 95% prædiktionsintervaller
pred <- predict(SLUTMODEL, newdata = D_test, interval = "prediction",
                level = 0.95)

# Observerede værdier sammen med prædiktioner
cbind(ETF = D_test$ETF, Geo.mean = D_test$Geo.mean, pred)
```

Dvs. skriv ikke formlerne ind i rapporten, men istedet, at I har brugt R funktionen `predict` til beregningerne. Formlerne kræver en matrix formulering, som rækker ud over pensum (for at udlede formlerne kan ligningerne (6-48) og (6-49) bruges sammen med udledningerne der fører til ligningerne (5-57) og (5-58)).