NTNU

Kunnskap for en bedre verden

Institutt for matematiske fag

# Eksamensoppgave i **TMA4267 Lineære statistiske modeller**

**Faglig kontakt under eksamen:** Mette Langaas

**Tlf:** 988 47 649

**Eksamensdato:** 4. juni 2016

**Eksamenstid (fra–til):** 09.00–13.00

**Hjelpemiddelkode/Tillatte hjelpemidler:** C: Gult, stemplet A5-ark med dine egne håndskrevne notater, Tabeller og formler i statistikk (Tapir forlag/Fagbokforlaget), K. Rottmann: Matematisk formelsamling. Bestemt kalkulator.

**Målform/språk:** bokmål

**Antall sider:** 10

**Antall sider vedlegg:** 0

**Kontrollert av:**

Dato Sign

## Oppgave 1 Uavhengige stokastiske variabler

La $\boldsymbol{X}$ være en bivariat normalfordelt stokastisk variabel med $\mathrm{E}(\boldsymbol{X}) = \begin{pmatrix} 5 \\ 3 \end{pmatrix}$ og $\mathrm{Cov}(\boldsymbol{X}) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$. La $\boldsymbol{Y} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \boldsymbol{X}$.

**a)** Finn fordelingen til $\boldsymbol{Y}$.

Spesifiser $a$, $b$ slik at $\boldsymbol{Y}$ og $\begin{pmatrix} 2 & a \\ b & 1 \end{pmatrix} \boldsymbol{X}$ er uavhengige stokastiske variabler. Begrunn svaret ditt.

## Oppgave 2 Plantestress

Forskere ved Institutt for biolog, NTNU, bruker modell-planten *Arabidopsis thaliana* for å studere hvordan en plante reagerer på ulike kilder til stress. I et forsøk ble planter utsatt for en stress-situasjon der følgende faktorer inngikk:

- $D$ (skade): $D = 1$ betyr at planten ble skadet mekanisk ved at bladene ble klippet med en saks. $D = -1$ betyr at ingen mekanisk skade ble påført (ingen klipping).
- $F$ (flagellin): $F = 1$ betyr at det patogen-deriverte peptidet flagellin ble sprayet på bladene til planten. $F = -1$ betyr at bare vann (ikke flagellin) ble sprayet på planten.
- $T$ (tid): Plantene ble høstet ved to ulike tidspunkt etter stress-situasjonen, $T = 1$ betyr at planten ble høstet 60 minutter etter stress-situasjonen, og $T = -1$ betyr at planten ble høstet 30 minutter etter stress-situasjonen.

Dermed inngår tre faktorer, $D$, $F$ and $T$, som hver kan ta to verdier. Forskerene utførte eksperimenter med alle mulige kombinasjoner av de tre faktorene fire ganger, slik at det totalt ble utført 32 eksperimenter.

Responsen som ble målt i dette eksperimentet var det observerte aktivitetsnivået (en kontinuerlig måling) til rundt 40 000 gener. Vi vil kun se på aktivitetsnivået til ett av disse genene, AT1G32920 genet, og vi kaller aktivitetsnivået til genet $Y$. Det er kjent at dette genet er aktivt ved skade på planten.

For eksperiment nummer $i$ (der $i = 1, \ldots, 32$): $Y_i$ er den observerte responsen, $D_i$ er den valgte verdien for $D$, $F_i$ er den valgte verdien for $F$, og $T_i$ er den valgte verdien til $T$. En multippel lineær regresjonsmodell med alle hovedeffekter, to- og tre-veis samspill, ble tilpasset

$$Y_i = \beta_0 + \beta_D D_i + \beta_F F_i + \beta_T T_i + \beta_{D:F} D_i F_i + \beta_{D:T} D_i T_i + \beta_{F:T} F_i T_i + \beta_{D:F:T} D_i F_i T_i + \varepsilon_i,$$

der $i = 1, \ldots, 32$, og vi antar at $\varepsilon_i$ er uavhengige og identisk normalfordelte med forventningsverdi 0 and varians $\sigma^2$. Vi kaller dette den *fulle modellen*. Merk at samspillene er produktene av faktorene. Vektoren med regresjonsparametere er $\boldsymbol{\beta} = \begin{pmatrix}\beta_0 & \beta_D & \beta_F & \beta_T & \beta_{D:F} & \beta_{D:T} & \beta_{F:T} & \beta_{D:F:T}\end{pmatrix}^{\mathrm{T}}$, og $i$te rad av designmatrisen $\boldsymbol{X}$ er $\begin{pmatrix}1 & D_i & F_i & T_i & D_iF_i & D_iT_i & F_iT_i & D_iF_iT_i\end{pmatrix}$.

I figur 1 finner du R-kommandoer og utskrift for den tilpassede fulle modellen.

**a)** I utskriften fra `summary(fit)` i figur 1 er *fire* tallverdier erstattet med spørsmålstegn. Regn ut tallverdier for hver av disse og forklar hva hvert av tallene betyr.

Et såkalt kubeplott er gitt i øvre venstre panel i figur 2. I kubeplottet finner du tilpassede verdier fra den multiple regresjonen for alle de 8 mulige kombinasjonene av de tre faktorene. Plott av hovedeffekter (øvre høyre panel) og samspillseffekter (nedre panel) finner du i figur 2. I figur 3 finner du residualplott. Se figur 4 for tilhørende R-kode og utskrift.

**b)** Hvordan vil du, fra figurene 2–4, vurdere modelltilpasningen?

Hvordan vil du forklare til en biolog hva den estimert hovedeffekten av skade betyr i praksis? Hvordan vil du forklare det estimerte samspillet mellom skade og flagellin?

La $\gamma = 2^{\beta_F - \beta_D}$ være en ny parameter som vi er interessert i.

Foreslå en estimator, $\hat{\gamma}$, for $\gamma$. Bruke tilnærmingsmetoder for å finne forventningsverdi og varians til denne estimatoren, det vil si, $\mathrm{E}(\hat{\gamma})$ og $\mathrm{Var}(\hat{\gamma})$. Bruk resultatene i figur 1 til å regne ut tallverdi for $\hat{\gamma}$, og estimerte tallverdier for $\mathrm{E}(\hat{\gamma})$ og $\mathrm{Var}(\hat{\gamma})$.

Hint: Du kan bruke at $2^x = \exp(x \ln 2)$, der ln er den naturlige logaritmen.

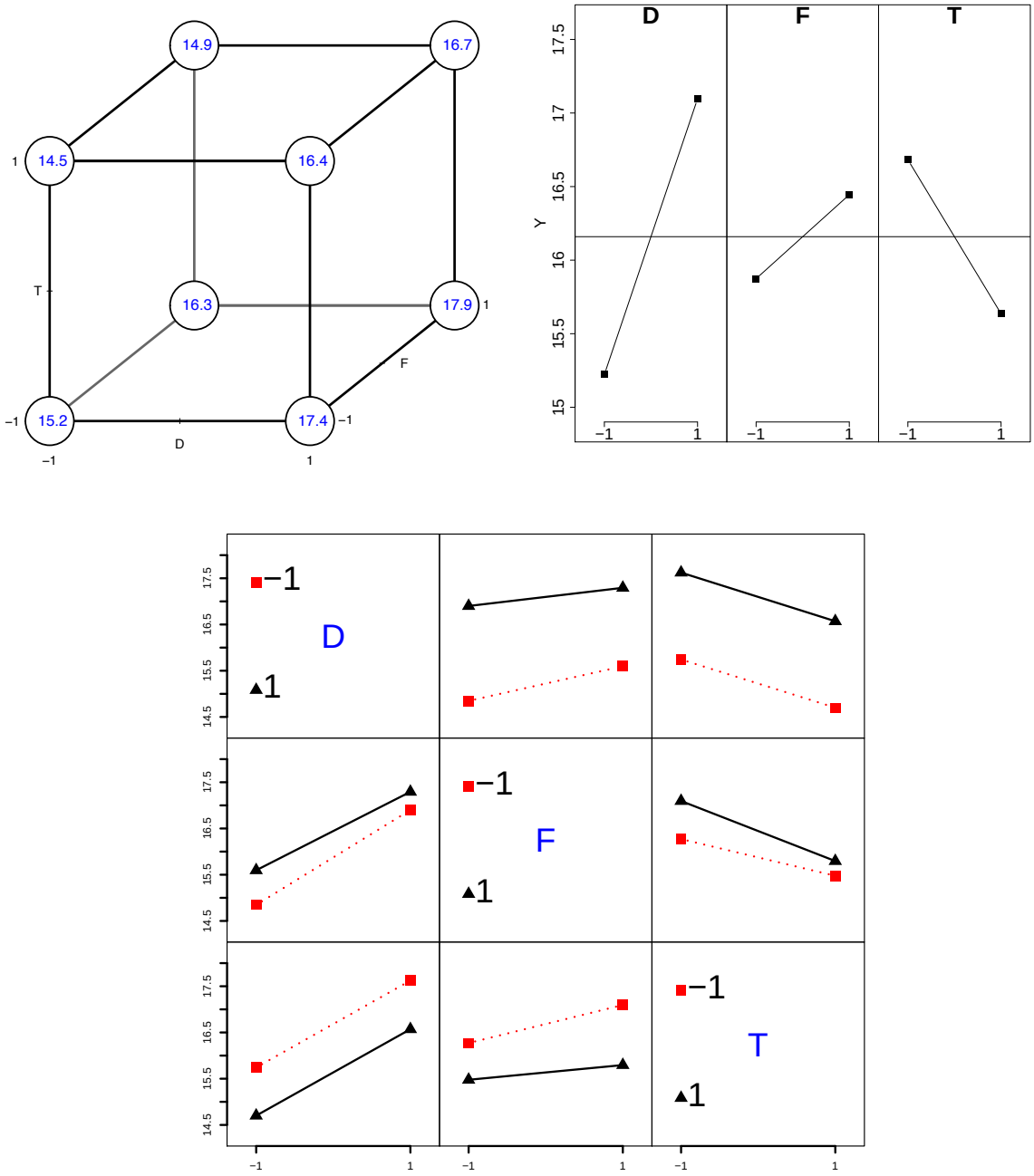
Forskerne vil teste hypotesen

$$H_0 \colon \beta_{D:T} = \beta_{F:T} = \beta_{D:F:T} = 0 \qquad \text{mot}$$
$$H_1 \colon \text{minst en } \beta_{D:T},\ \beta_{F:T},\ \beta_{D:F:T} \text{ er ulik 0.}$$

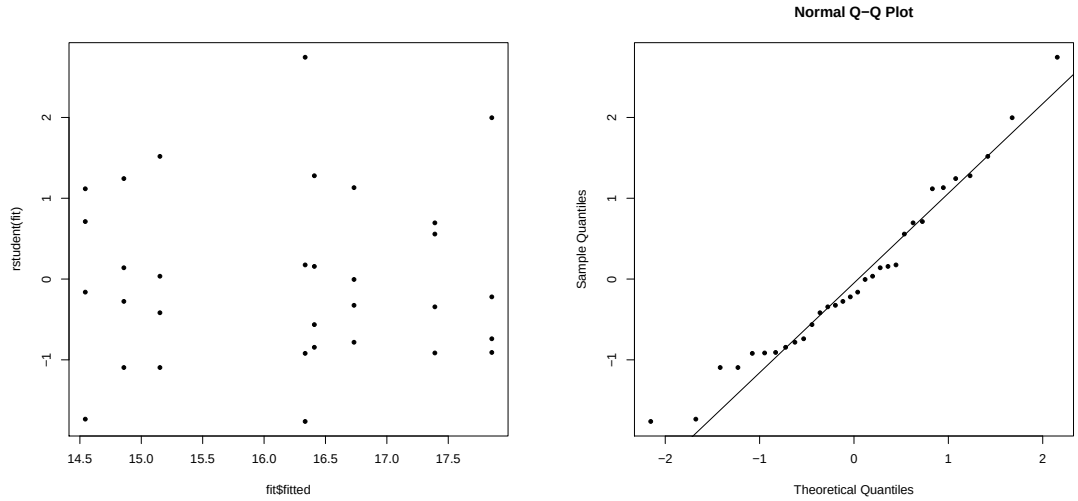
**c)** Velg signifikansnivå selv og utfør hypotesetesten. Alle tallverdier du trenger for beregningene finner du i figur 1.

```
# data is in "standard order" in data frame with name "ds"
> ds %showing only rows 1-6 and 27-32 for space considerations
          Y  D  F  T
1  15.45169 -1 -1 -1
2  15.15908 -1 -1 -1
3  14.93064 -1 -1 -1
4  15.06569 -1 -1 -1
5  14.51032 -1 -1  1
6  14.76922 -1 -1  1
...
27 18.23645  1  1 -1
28 17.70327  1  1 -1
29 16.66523  1  1  1
30 16.96046  1  1  1
31 16.73133  1  1  1
32 16.57248  1  1  1
> fit=lm(Y~D*F*T,data=ds)
> model.matrix(fit)%only showing rows 1-6 and 27-32
   (Intercept)  D  F  T D:F D:T F:T D:F:T
1            1 -1 -1 -1   1   1   1    -1
2            1 -1 -1 -1   1   1   1    -1
3            1 -1 -1 -1   1   1   1    -1
4            1 -1 -1 -1   1   1   1    -1
5            1 -1 -1  1   1  -1  -1     1
6            1 -1 -1  1   1  -1  -1     1
...
27           1  1  1 -1   1  -1  -1    -1
28           1  1  1 -1   1  -1  -1    -1
29           1  1  1  1   1   1   1     1
30           1  1  1  1   1   1   1     1
31           1  1  1  1   1   1   1     1
32           1  1  1  1   1   1   1     1
> summary(fit)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.15942    0.04140  ?       < 2e-16
D            0.93739    0.04140  22.644  < 2e-16
F            0.28546    0.04140   6.896 3.93e-07
T           -0.52354    0.04140 -12.647 4.18e-12
D:F         -0.08878    0.04140  -2.145  0.04231
D:T         -0.00242    ?        -0.058  0.95386
F:T         -0.12614    0.04140  -3.047  0.00555
D:F:T        0.09099    0.04140   2.198  ?
Residual standard error: 0.2342 on 24 degrees of freedom
Multiple R-squared:      ?, Adjusted R-squared:  0.9594
F-statistic: 105.6 on 7 and 24 DF,  p-value: < 2.2e-16
```

Figur 1: Utskrift av R-kommandoer og statistisk analyse for plantestress-datasettet. Fire tallverdier er erstattet med spørsmålstegn.




Figur 2: Kubeplott (øvre venstre panel), hovedeffektsplott (øvre høyre panel) og interaksjonseffektplott (nedre panel) for den fulle regresjonmodellen tilpasset til plantestress-datasettet.




Figur 3: Residualplott (studentiserte residualer mot tilpassede verdier i venstre panel og normalplott basert på studentiserte residualer i høyre panel) for den fulle modellen tilpasset til plantestress-datasettet.

```
> library(FrF2)
> MEPlot(fit)
> IAPlot(fit)
> cubePlot(fit,"D","F","T",round=1,size=0.33,main="")
> plot(fit$fitted,rstudent(fit),pch=20)
> qqnorm(rstudent(fit),pch=20)
> qqline(rstudent(fit))
> ad.test(rstudent(fit))
        Anderson-Darling normality test
data:  rstudent(fit)
A = 0.43191, p-value = 0.2869
```

Figur 4: Utskrift av R-kommandoer og statistisk analyse for den fulle modellen for plantestress-datasettet.

Forskerne ønsker å bruke dataene til prediksjon, og vil tilpasse en redusert versjon av den fulle modellen. Først ble «best subset»-metoden brukt. Deretter tilpasset forskerene en lasso-regresjon til dataene. Resultater finner du i figur 5 og 6.

**d)** Forklar kort hva som gjøres i «best subset»-metoden, og velg en god modell basert på $R^2_{\text{adj}}$-kriteriet. Skriv ned den tilpassede regresjonsmodellen du velger.

Forklar kort hva som gjøres i lasso-regresjonen, og skriv ned den tilpassede regresjonsmodellen.

Sammenlign resultatene fra «best subset»-regresjonen og lasso-regresjonen.

Forskerene velger å bruke følgende *reduserte modell* for prediksjon:

$$Y_i = \beta_0 + \beta_D D_i + \beta_F F_i + \beta_T T_i + \beta_{D:F} D_i F_i + \varepsilon_i,$$

der $i = 1, \ldots, 32$, og vi antar at $\varepsilon_i$ er uavhengige og identisk normalfordelte med forventningsverdi 0 and varians $\sigma^2$. Utskrift fra den tilpassede reduserte modellen finner du i figur 7.

**e)** Sammenlign de estimerte regresjonsparameterene og de estimert standardavvikene til de estimerte regresjonsparameterene for den fulle modellen (figur 1) og den reduserte modellen (figur 7), og forklar hva du observerer.

Bruk den reduserte modellen (figur 7) til å lage en prediksjon og et 95% prediksjonsintervall for genaktivitetsnivået for faktorkombinasjonen $D = 1$, $F = 1,\ T = -1$.

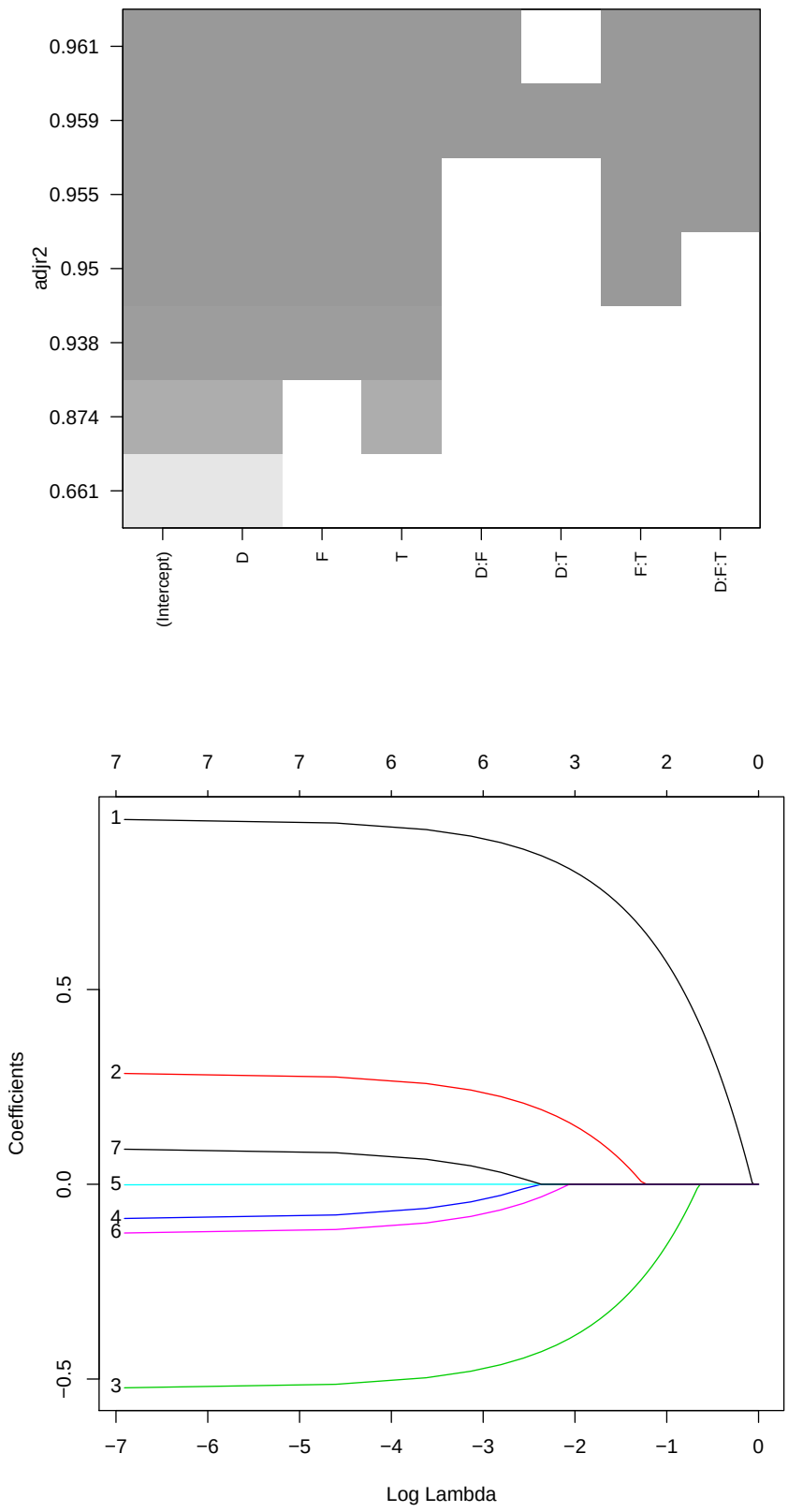
Hint: I en multippel lineær regresjon med $n \times p$ designmatrise $\boldsymbol{X}$, estimerte regresjonsparametere $\hat{\boldsymbol{\beta}}$ og forventningsrett estimert feilvarians $s^2$, er et $(1-\alpha)100\%$ prediksjonsintervall i $\boldsymbol{x}_0$ gitt ved

$$\boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} \pm t_{\frac{\alpha}{2},n-p}\, s \sqrt{1 + \boldsymbol{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_0},$$

der $t_{\alpha/2,n-p}$ er verdien i $t$-fordelingen med $n-p$ frihetsgrader som har areal $\frac{\alpha}{2}$ til høyre. Noen mulige verdier for $t_{\alpha/2,n-p}$ er gitt i figur 7.

```
> x <- model.matrix(fit)[,-1]; dim(x)
[1] 32  7
> y <- ds$Y
> library(leaps)
> bests <- regsubsets(x,y)
> sumbests=summary(bests)
> sumbests
1 subsets of each size up to 7
Selection Algorithm: exhaustive
         D   F   T   D:F D:T F:T D:F:T
1  ( 1 ) "*" " " " " " " " " " " " " "
2  ( 1 ) "*" " " "*" " " " " " " " " "
3  ( 1 ) "*" "*" "*" " " " " " " " " "
4  ( 1 ) "*" "*" "*" " " " " "*" " " "
5  ( 1 ) "*" "*" "*" " " " " "*" "*"
6  ( 1 ) "*" "*" "*" "*" " " "*" "*"
7  ( 1 ) "*" "*" "*" "*" "*" "*" "*"
> plot(bests,scale="adjr2",col=gray(seq(0.6,0.9,length=20)))
> round(sumbests$adjr2,3)
[1] 0.661 0.874 0.938 0.950 0.955 0.961 0.959
# LASSO
> library(glmnet)
> fit.lasso=glmnet(x,y,lambda=c(seq(1,0.01,length=60),0.001))
> plot(fit.lasso,xvar="lambda",label=TRUE)
> cv.lasso=cv.glmnet(x,y)
> log(cv.lasso$lambda[which.min(cv.lasso$cvm)])
[1] -4.716347
> coef(cv.lasso,s="lambda.min")
8 x 1 sparse Matrix of class "dgCMatrix"
(Intercept) 16.15941869
D            0.92843876
F            0.27651524
T           -0.51459006
D:F         -0.07983558
D:T          .
F:T         -0.11719275
D:F:T        0.08204094
```

Figur 5: Utskrift fra R for «best subset»-modellvalg og lasso-regresjon for plantestress-datasettet.




Figur 6: Figur (øvre panel) fra «best subset»-modellseleksjonen og lasso-regresjonen (nedre panel), med R-kode i figur 5. Koder for linjene i lasso-figuren er 1=$\hat{\beta}_D$, 2=$\hat{\beta}_F$, 3=$\hat{\beta}_T$, 4=$\hat{\beta}_{D:F}$, 5=$\hat{\beta}_{D:T}$, 6= $\hat{\beta}_{F:T}$ og 7=$\hat{\beta}_{D:F:T}$.

```
> fitRED=lm(Y~D+F+T+D:F,data=ds)
> summary(fitRED)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.15942    0.04919 328.528  < 2e-16
D            0.93739    0.04919  19.057  < 2e-16
F            0.28546    0.04919   5.804 3.56e-06
T           -0.52354    0.04919 -10.644 3.66e-11
D:F         -0.08878    0.04919  -1.805   0.0822
Residual standard error: 0.2782 on 27 degrees of freedom
Multiple R-squared:   0.95,        Adjusted R-squared:  0.9426
F-statistic: 128.4 on 4 and 27 DF,  p-value: < 2.2e-16
> qt(0.025,32,lower.tail=FALSE)
[1] 2.036933
> qt(0.025,27,lower.tail=FALSE)
[1] 2.051831
> qt(0.025,24,lower.tail=FALSE)
[1] 2.063899
```

Figur 7: Utskrift fra R for multippel lineær regresjon basert på den reduserte modellen for plantestress-datasettet.

## Oppgave 3 Egenskaper til estimator for $\sigma^2$

La $\boldsymbol{Y}$ være en $n \times 1$ stokastisk vektor med forventningsverdi $\mu\mathbf{1}$ og kovariansmatrise $\sigma^2\boldsymbol{I}$, der $\mathbf{1}$ er en $n \times 1$ vektor med alle elementer lik 1 og $\boldsymbol{I}$ er en $n \times n$ identitetsmatrise. Videre er $Y_i$ element $i$ fra $\boldsymbol{Y}$, og $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\mathbf{1}^{\mathrm{T}}\boldsymbol{Y}$.

En estimator for $\sigma^2$ er

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \frac{1}{n-1}\boldsymbol{Y}^{\mathrm{T}}\Big(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}\Big)\boldsymbol{Y}.$$

Vi oppgir følgende nyttige resultat. La $\boldsymbol{X}$ være en $n \times 1$ stokastisk vektor med forventningsverdi $\boldsymbol{\eta}$ og kovariansmatrise $\boldsymbol{\Sigma}$, og la $\boldsymbol{C}$ være en $n \times n$ symmetrisk konstantmatrise. Da er

$$\mathrm{E}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{X}) = \mathrm{tr}(\boldsymbol{C}\boldsymbol{\Sigma}) + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{C}\boldsymbol{\eta}. \tag{1}$$

**a)** Skriv først ned verdien til $\mathbf{1}^{\mathrm{T}}\mathbf{1}$, og til matrisene $\mathbf{1}\mathbf{1}^{\mathrm{T}}$ og $\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}$ for $n = 4$.
Hva er nøkkelegenskaper til matrisen $\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}}$ (symmetrisk eller ikke, idempotent eller ikke, rang)?
Bruk ligning (1) til å finne $\mathrm{E}(S^2)$.

La oss anta at $\boldsymbol{Y}$ er multivariat normalfordelt med forventningsverdi og kovariansmatrise gitt over.

**b)** Vis at $\frac{1}{\sigma^2}\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y}$ er $\chi^2$-fordelt, og finn antall frihetsgrader.
Bruk dette resultatet til å finne variansen til $S^2$.
Er den stokastiske variabelen $\frac{1}{n}\mathbf{1}^{\mathrm{T}}\boldsymbol{Y}$ og den stokastiske vektoren $(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y}$ uavhengige? Begrunn svaret.
Finn til slutt fordelingen til

$$\frac{n(\frac{1}{n}\mathbf{1}^{\mathrm{T}}\boldsymbol{Y} - \mu)^2}{\frac{1}{n-1}\boldsymbol{Y}^{\mathrm{T}}(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^{\mathrm{T}})\boldsymbol{Y}}.$$

Begrunn svaret.