EKSTRAOPPGAVER I STK1110 H2018

1. Simuleringer for å illustrere store talls lov og sentralgrenseteoremet

Oppgave 1.1. I denne oppgaven skal vi bruke kommandoen `rbinom(n,size,prob)`. Kommandoen trekker `n` tilfeldige variable fra binomisk fordeling med parametre `size` og `prob`. Spesielt for `size=1` trekkes `n` uavhengige variable med mulige utfall lik 0 eller 1, der sannsynligheten for å trekke 1 er lik `prob`. Eksempel: Kommandoen `rbinom(n=10,size=1,prob=0.5)` kan gi oss tallene: 1 1 0 0 0 1 0 1 1 0. Her ser vi at vi fikk 10 tall med mulige utfall lik 0 eller 1, med sannsynlighet lik 0.5 for å få 1. Siden vi her trekker tall vil resultatene og plasseringen av enerne variere fra gang til gang.

Vi tenker oss nå at vi kaster en mynt 25 ganger, og noterer antall kron (=1) og mynt (=0) der sannsynligheten for kron er $p = 0.6$. Vi gjentar trekningene 100 ganger, og lager et histogram over andelene $\hat{p}$ av kron i disse 100 forsøkene. Du kan skrive:

```
hist(rbinom(n=100,size=25,prob=0.6)/25)
```

Beskriv fordelingen for estimatene for sannsynligheten $p = 0.6$. Hvor er denne sentrert? Beskriv variasjonen i $\hat{p}$. Ligner histogrammet på en normalfordeling?

Oppgave 1.2. Oppgaven er tilsvarende oppgaven over, men vi er nå interessert i å se på hvordan $\hat{p}$ utvikler seg etterhvert som man kaster mynten flere ganger. Da kan antall og andelene i hvert av de 100 forsøkene på $n = 50$ myntkast med sannsynlighet lik 0.6 for kron genereres ved

```
ant=rbinom(n=100,size=50,prob=0.6)
phat=ant/50
```

Lag et histogram `hist(phat)` over $\hat{p}$-ene. Det kan også være nyttig å beskrive fordelingen ved `summary(phat)`.

Gjenta trekningene med andre verdier for `size`. Endres senteret i histogrammet nevneverdig når `size` endres, mens `prob` holdes konstant? Hva skjer med variasjonen og normaltilnærmingen til $\hat{p}$-ene?

Velg andre verdier av `prob`, f.eks. 0.1 og 0.9. Hvor ligger senteret i histogrammet. Hvordan endres fordelingen nå når n gjøres større?

Oppgave 1.3. (a) Det skal trekkes en serie på uavhengige 0-1 variable med sannsynlighet 0.5 for begge verdier. Vi skal bå de se på hvordan *andelen* av 1-ere utvikler seg (og stabiliserer seg mot 0.5), men i tillegg hvordan *antallet* 1-ere utvikler seg ettersom man trekker flere verdier. For å se dette bedre vil vi trekke fra forventet antall 1-ere lik $\frac{n}{2}$.

Vi skal bruke R-kommandoen `cumsum`. Navnet er kort for kumulativ (oppsamlende) sum og fra en vektor $(x_1, x_2, x_3, \ldots, x_n)$ genererer den vektoren med elementene $(x_1, x_1 + x_2, x_1 + x_2 + x_3, \ldots, x_1 + x_2 + \cdots + x_{n-1} + x_n)$. Vi får altså summene av de første k verdiene for $k = 1, 2, \ldots, n$.

```
n=50
x=rbinom(n,1,0.5)
kumsumx=cumsum(x)
phat=kumsumx/(1:n)
plot(1:n,phat,type="l",xlab="Antall myntkast",ylab="Andel suksesser",ylim=c(0,1))
abline(h=0.5,lty=3)
plot(1:n,kumsumx-0.5*(1:n),type="l",ylab="Totalt avvik fra forventning")
abline(h=0,lty=3)
```

Bytt så ut $n = 50$ med $n = 150$ og $n = 300, 600, \ldots$ (du kan godt gå opp til en million).

Poenget med oppgaven er å se at det er andelen (gjennomsnittet) som stabiliserer seg når n vokser, mens antallet vil avvike stadig mer fra forventningen.

Oppgave 1.4. (a) Vi skal her simulere lange serier av terningkast med den hensikt å illustrere *store talls lov*. Dette resultatet sier at gjennomsnittet av uavhengige tilfeldige variable vil gå mot forventningen μ. For den tilfeldige variablen gitt ved antall øyne på en terning er forventningen $\mu = 3.5$.

```
n=100
antoyne=sample(1:6,n,replace=T) #replace=T siden vi skal trekke med tilbakelegging
kumoyne=cumsum(antoyne)
gjsnoyne=kumoyne/(1:n)
plot(1:n,gjsnoyne,type="l",xlab="Antall kast",ylab="Gjennomsnitt",ylim=c(1,6))
abline(h=3.5,lty=3)
```

Beskriv utviklingen i gjennomsnittelig antall øyne når antall kast vokser og sammenhold med forventningsverdien $\mu = 3.5$.

Bytt (gjerne) ut $n = 100$ med noen passende større antall kast!

Oppgave 1.5. Under generelle betingelser vil summen av tilfeldige observasjoner være normalfordelt. Dette er kjent som Sentralgrenseteoremet og er ofte alternativt formulert som at gjennomsnittet $\frac{1}{n}\sum X_i$ er tilnærmet normalfordelt.

Det er stort sett dette matematiske resultatet vi lener oss på når vi gjør statistiske tester på datasett som faktisk *ikke er normalfordelt*; ofte vil betingelsene til sentralgrenseteoremet være oppfylt, og man kan så vise at testene er tilnærmet riktige likevel. Det er til tross for at vi ikke kjenner fordelingen til datasettet.

Dette er overraskende, og slett ikke opplagt; hvorfor skulle summer oppføre seg så annerledes enn hvert av leddene? Man kan, ved hjelp av en del matematikk, vise dette resultatet. Vi, derimot, skal i denne oppgaven se at dette fungerer ved hjelp av et *simuleringsforsøk*.

(a) Først skal vi illustrere at summen av tilfeldige variable er tilnærmet normalfordelt for uniforme variable beskrevet. Trekk først 1000 uniforme variable ved kommandoen `x=runif(1000)`. Sjekk deretter tettheten til uniforme variable ved å tegne et histogram (`hist(x)` og verifiser at denne er langt fra normal med kommandoene `qqnorm(x)` og `qqline(x)`.

(b) Se så på summer av to uniforme variable og lag histogram og kvantilplott som i punkt (a). Du genererer disse summene ved å skrive `x=runif(1000)+runif(1000)`. Se hvordan tettheten til summen av de to uniforme variablene er, og sammenlign med resultatet i (a).

(c) Se så på summen av $n = 12$ uniforme variable og vurder tilnærming til normalfordeling ved å se på histogram og kvantilplott for de simulerte summene. For å simulere disse summene benytter du funksjonen `replicate(M,f)`, som M ganger gjentar funksjonen `f` For å generere $M = 1000$ summer av $n = 12$ uniforme variable kan du da skrive:

```
n = 12
SumX = replicate(1000,sum(runif(n)))
hist(SumX)
qqnorm(SumX)
qqline(SumX)
```

(d) Gjenta gjerne simuleringen med en ny (mindre!) n og gjør en vurdering av hvor stor denne må være før man har en tilfredstillende tilnærming til normalfordelingen for uniforme tilfeldige variable.

(e) Grunnen til at vi for uniforme variable trenger så liten n for at $X_1+\cdots+X_n$ skal være godt tilnærmet med en normalfordeling er at den uniforme fordelingen er symmetrisk og har (ekstremt) lette haler. Det vil dermed kreves et større antall n når fordelingen til X_i-ene er skjev med tunge haler. En slik fordeling er eksponensialfordelingen. Vi skal i det følgende replisere dette eksempelet med simulerte data. Først trekk 1000 eksponensielle variable ved `x=rexp(1000)` og lag histogram og kvantilplott som over samt gjerne også boksplott.

(f) Trekk så 1000 summer av to eksponensielle variable ved `x=rexp(1000)+rexp(1000)` og vurder tetthet og tilnærming til normalfordeling.

(g) Gjenta dette med f.eks. $n = 10, 25$ og 100. Med $n > 2$ benytter du `R`-funksjonen `replicate` beskrevet i (c) og bytter ut `runif(n)` med `rexp(n)`.

(h) Prøv gjerne med å trekke fra andre fordelinger enn eksponensial og uniform fordelingene.

2. Konvolusjonsformelen med andvendelser

Oppgave 2.1. I STK1100 brukte man momentgenerende funksjoner (mgf) for å finne fordelingen til summen $X + Y$ av to uavhengige stokastiske variable X og Ymed kjente tettheter $f(x)$ og $g(y)$. Dette er ofte det matematisk mest bekvemme, men fungerer bare når mgf-ene for X og Y er kjente og når man gjenkjenner fordelingen svarende til produktet av mgf-ene.

En generell metode for å finne fordelingen til $Z = X + Y$ gis ved konolvusjonsformelen

$$h(z) = \int f(x)g(z-x)dx$$

der integralet er over alle mulige verdier av x. (En enkel modifikasjon av denne formelen tillater faktisk også avhengige X og Y med simultanfordeling $f(x, y)$: $h(z) = \int f(x, z-x)dx$.)

(a) Vi skal ikke utlede dette konvolusjonsformelen helt generelt, men vise analogien når X og Y er diskrete stokastiske variable på heltallene $0, 1, \ldots, n$ med puktsannsynligheter $f(x)$ og $g(y)$. Da gjelder følgende formel, som ligner på konvolusjonsformelen:

$$P(Z = X + Y = z) = \sum_{x=0}^{n} f(x)g(z-x).$$

Vis at denne formelen holder.

Hint: Bruk loven om total sannsynlighet, formel (2.5) på side 79 i Devore & Berk.

(b) Anta $X \sim Bin(n_1, p)$ og $Y \sim Bin(n_2, p)$ er uavhengige. Vis ved bruk av momentgenerende funksjoner at $Z = X + Y \sim Bin(n_1 + n_2, p)$.

Verifiser resultatet ved å bruke konvolusjonsformelen i punkt a).

Hint: Forklar hvorfor og bruk at $\sum_x \binom{n_1}{x}\binom{n_2}{z-x} = \binom{n_1+n_2}{z}$ når summen går over x slik at $\max(0, n_2 - z) \leq x \leq \min(n_1, z)$.

(c) Anta nå at X og Y er uavhengige og Poissonfordelte med forventninger hhv. λ og μ. Gjenta argumenter fra STK1100 for å påvise at $X + Y$ er Poissonfordelt med forventning $\lambda + \mu$.

Benytt alternativt konvolusjonsformelen fra punkt a) til å utlede det samme resultatet.

Hint: Benytt binomialformelen som kan skrives $(\lambda + \mu)^z = \sum_{x=0}^{z} \binom{z}{x}\lambda^x\mu^{z-x}$.

(d) Benytt momentgenerende funksjon til å vise at hvis X og Y er uavhengige og gammafordelte med samme skalaparameter β og formparametre hhv. α_1 og α_2 så er $X + Y$ gammafordelt med parametre $\alpha_1 + \alpha_2$ og β.

Vis deretter først at hvis $\alpha_1 = \alpha_2 = 1$, dvs. når X og Y er eksponensialfordelte med forventning β så er $X + Y$ gammafordelt med parametre $\alpha = 2$ og β ved å bruke konvolusjonsformelen for kontinuerlige fordelte stokastiske variable.

Utvid så resultatet til generelle gammafordelinger.

Hint: Husk tettheten til beta-fordelingen som i standardform skrives som $g(x; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$ for $0 < x < 1$. Muligens er det en hjelp å bruke den generelle versjonen av beta-fordelingen på side 207 i Devore & Berk.

(e) Endelig, anta at X og Y er uavhengige og standardnormalfordelte. Da er som kjent $Z = X + Y$ normalfordelt med forventning $\mu = 0$ og varians $\sigma^2 = 2$. Vis dette resultatet ved hjelp av konvolusjonsformelen.

Hint: Lag fullstendig kvadrat.

3. Estimering, konfidensintervaller og hypotesetesting

Oppgave 3.1. I denne oppgave skal du sammenligne moment og maximum likelihood estimatorer basert på uif variable $X_i, i = 1, \ldots, n$ med tetthet $f_X(x;\theta)$ med de tilsvarende estimatorene for $Y_i = g(X_i)$ der $g()$ er en en-entydig og deriverbar funksjon.

(a) La $h(y)$ være den inverse funksjonen til $g()$. Forklar hvorfor Y_i-ene er uif med tetthet $f_Y(y;\theta) = f_X(h(y);\theta)h'(y)$.

La $L_X(\theta)$ være likelihooden basert på X_i-ene og $L_Y(\theta)$ likelihooden basert på Y_i-ene. Vis at disse to likelihoodene er proporsjonale. Hva sier dette om sammenhengen mellom maximum likelihood estimatorer basert på X_i-er og på Y_i-er?

(b) Anta at $X_i \sim \mathrm{N}(\mu, \sigma^2)$. Vis at momentestimatorene for μ og σ^2 gis ved henholdsvis $\hat{\mu} = \bar{X}$ og $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$.

(c) Vis at maximum likelihoodestimatorene for μ og σ^2 i dette tilfellet er identiske med momentestimatorene.

(d) La $Y_i = \exp(X_i)$ slik at Y_i-ene er log-normalfordelte med parametre μ og σ^2. Hva blir maximum likelihood estimatorene for μ og σ^2?

(e) Den momentgenerende funksjonen til X_i-ene kan skrives $M_X(t) = \mathrm{E}[\exp(tX)] = \exp(\mu t + \sigma^2 t^2/2)$. Bruk dette til å bestemme at $\mathrm{E}[Y_i] = \exp(\mu + \sigma^2/2)$ og $\mathrm{E}[Y_i^2] = \exp(2\mu + 2\sigma^2)$.

Vis så at momentestimatoren for σ^2 kan skrives $\tilde{\sigma}^2 = \ln(\frac{1}{n}\sum_{i=1}^n Y_i^2) - 2\ln(\bar{Y})$.

Hva blir momentestimatoren for μ basert på Y_i-ene?

(f) Simuler $n = 10$ verdier av $X_i \sim \mathrm{N}(1, 3)$ ved kommandoen `x=rnorm(10,1,sqrt(3))`. Beregn momentestimatorene for (μ, σ^2) basert på X_i-er og Y_i-er og sammenlign. Er de to momentestimatorene forskjellige?

(g) Gjenta simuleringene 1000 ganger og beregn gjennomsnitt og empiriske varianser over simuleringene. Hva kan du si om skjevheter og varianser for de ulike estimatorene?

(h) Gjenta simuleringer som over, men nå med $n = 1000$ observasjoner i hver simulering. Sammenlign igjen estimatorene. Hvordan endres skjevheter og varianser?

(i) Se videre på histogram, boksplott og qq-plott for estimatene - både i punkt (g) og punkt (h). Kommenter om normaltilnærming og outliere.

(j) Se gjerne også på hva som skjer hvis man velger andre verdier av μ og σ^2 i simuleringene. Finner du noe interessant?

Oppgave 3.2. Vi skal i denne oppgaven se nærmere på data fra oppgave 8.10 i læreboka. Her er et tilfeldig utvalg av $n = 15$ varmepumper undersøkt mhp levetid, noe som ga følgende levetider (i år):

```
2.0  1.3  6.0  1.9  5.1  0.4  1.0  5.3 15.7  0.7  4.8  0.9 12.2  5.3  0.6
```

Vi vil anta at levetidene er eksponensielt fordelt med parameter λ slik at

$$f(x;\lambda) = \lambda e^{-\lambda x}, \quad x \geq 0$$

Vi ønsker å teste hypotesen

$$H_0 : \lambda = 0.35 \text{ mot } H_a : \lambda \neq 0.35 \qquad (*)$$

(a) Finn maksimum likelihood estimatet for λ.

(b) Vis at hvis X_i er eksponensielt fordelt, så er $2\lambda X_i$ kjikvadrat fordelt med 2 frihetsgrader.

Vink: Regn ut $E[e^{2t\lambda X}]$ og relater dette til momentgererende funksjoner.

(c) Vis at $2\lambda \sum_{i=1}^{n} X_i$ er kjikvadrat fordelt med $2n$ frihetsgrader og bruk dette til å konstruere et konfidensintervall for λ.

(d) En generell sammenheng mellom konfidensintervaller og hypotesetesting gir at nullhypotesen $\mathrm{H}_0 : \lambda = \lambda_0$ mot alternativet $\mathrm{H}_0 : \lambda \neq \lambda_0$ forkastes med nivå α hvis λ_0 ikke er med i konfidensintervallet med konfidensgrad $(1-\alpha)100\%$.

Utfør en test basert på sammenhengen mellom konfidensintervall og hypotesetesting.

Hva blir din konklusjon på testen hvis du bruker $\alpha = 0.05$?

Hva blir din konklusjon på testen hvis du bruker $\alpha = 0.1$?

(e) Argumenter for at P-verdien til denne testen ligger mellom 0.05 og 0.1.

Hva blir P-verdien basert på denne testen?

Vink: Bruk Proposisjon på side 458 i boka.

(f) Konstruer nå en Likelihood ratio test for å teste hypotesen. Likelihood ratio for hypotesetesten i punkt d) er gitt ved $\mathrm{LR} = L(\lambda_0)/L(\hat{\lambda})$ der $L(\lambda)$ er likelihood med parameter λ og $\hat{\lambda}$ er maximum likelihood estimatoren for λ. Vis at

$$-2\ln(\mathrm{LR}) = 2n\ln(\hat{\lambda}) - 2n\ln(\lambda_0) - 2(\hat{\lambda} - \lambda_0)\sum_{i=1}^{n} x_i$$

Bruk dette til å teste (*). Hva blir konklusjonen hvis du bruker $\alpha = 0.05$ og tilsvarende når du bruker $\alpha = 0.1$.

Vink: Bruk de generelle egenskaper om Likelihood ratio observatorer som beskrevet på side 477 i boka.

(g) Beregn P-verdien for LR-testen.

(h) Diskuter eventuelle forskjeller mellom testen basert på konfidensintervaller og testen basert på LR.

Oppgave 3.3. Anta $X_1, ..., X_n \overset{uif}{\sim} N(\mu, \sigma)$ der σ er kjent. Vi ønsker å teste

$$H_0 : \mu = \mu_0 \quad \text{mot } H_a : \mu \neq \mu_0$$

(a) Formuler hypotesene som

$$H_0 : \theta \in \Omega_0 \quad \text{mot } H_a : \theta \in \Omega_a$$

Hva blir $\Omega = \Omega_0 \cup \Omega_a$ i dette tilfellet?

(b) Vis at under H_0, er

$$L(\widehat{\Omega}_0) = \max_{\theta \in \Omega_0} L(\theta) = \left(\tfrac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu_0)^2}$$

(c) Vis at

$$L(\widehat{\Omega}) = \max_{\theta \in \Omega} L(\theta) = \left(\tfrac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \bar{x})^2}$$

(d) Finn

$$\text{LR} = \tfrac{L(\widehat{\Omega}_0)}{L(\widehat{\Omega})}$$

og vis at LR er liten er ekvivalent med at $|Z|$ er stor der $Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$.

(e) Vis at $-2\ln(\text{LR})$ er χ_1-fordelt *eksakt* i dette tilfellet.

Oppgave 3.4. Anta $X_1, ..., X_n \overset{uif}{\sim}$ Geometrisk(p), dvs den geometriske sannsynlighetsfordeling med suksess-sannsynlighet p. Da er

$$\Pr(X = k) = (1-p)^{k-1} p, \quad k = 1, 2, ...$$

Vi ønsker å teste

$$H_0 : p = p_0 \quad \text{mot } H_a : p \neq p_0$$

(a) Vis at Maksimum likelihood estimatet for p er

$$\hat{p} = \tfrac{n}{\sum_{i=1}^n x_i}$$

(b) Vis at Likelihood ratio testen blir

$$\text{LR} = \left(\tfrac{1-p_0}{1-\hat{p}}\right)^{\sum_{i=1}^n x_i - n} \left(\tfrac{p_0}{\hat{p}}\right)^n$$

(c) Anta vi har observert

```
1  6  8  2 12  1 13  8  1  2  2  3  6  3 12  2  2  3 15  7
```

Hva blir $\hat{p}$ i dette tilfellet?

(d) Utfør en LR-test med $p_0 = 0.2$ og $\alpha = 0.05$. Hva blir konklusjonen av testen?

4. Toutvalgsdata, Kap. 10

Oppgave 4.1. Vi skal i denne oppgaven se på F-fordelingen og p-verdier for to-sidige F-tester. Vi har altså at hvis $Z_1 \sim \chi^2_{\nu_1}, Z_2 \sim \chi^2_{\nu_2}$ og Z_1 og Z_2 er uavhengige, så er

$$F = \frac{Z_1/\nu_1}{Z_2/\nu_2}$$

F-fordelt med ν_1 og ν_2 frihetsgrader.

(a) Vis at $1/F$ også er F-fordelt, men med ν_2 og ν_1 frihetsgrader

(b) Vis at hvis F_{α,ν_1,ν_2} er øvre α kvantil i F-fordelingen med ν_1 og ν_2 frihetsgrader, så er

$$F_{1-\alpha,\nu_1,\nu_2} = \frac{1}{F_{\alpha,\nu_2,\nu_1}}$$

Anta nå $X_1, ..., X_m \overset{uif}{\sim} N(\mu_1, \sigma_1^2)$ og $Y_1, ..., Y_n \overset{uif}{\sim} N(\mu_2, \sigma_2^2)$ samt at X-ene og Y-ene er uavhengige. Vi vil teste hypotesene

$$H_0 : \sigma_1^2 = \sigma_2^2 \text{ mot } H_a : \sigma_1^2 \neq \sigma_2^2$$

og vil forkaste H_0 hvis

$$f = \frac{s_1^2}{s_2^2}$$

enten er større enn $F_{\alpha/2;m-1,n-1}$ eller mindre enn $F_{1-\alpha/2;m-1,n-1}$

(c) Vis at

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

er F fordelt med $m-1$ og $n-1$ frihetsgrader og bruk dette til å vise at signifikansnivået for testen er α.

(d) For et ensidig alternativ $H_a : \sigma_1^2 > \sigma_2^2$, argumenter for hvorfor $F > f$ er en mer ekstrem verdi og tilsvarende at for et ensidig alternativ $H_a : \sigma_1^2 < \sigma_2^2$ så er $F < f$ en mer ekstrem verdi.

Bruk dette til å argumentere for at vi kan beregne en P-verdi for en to-sidig test ved

$$\text{P-verdi} = \begin{cases} 2\Pr(F > f) & \text{hvis } f > 1 \\ 2\Pr(F < f) & \text{hvis } f < 1 \end{cases}$$

5. Regresjon og korrelasjon, Kap. 12

Oppgave 5.1. Anta at (X, Y) er bivariat normalfordelt med forventninger $\mu_1 = \mathrm{E}[X]$ og $\mu_2 = \mathrm{E}[Y]$, varianser $\sigma_1^2 = \mathrm{V}[X]$ og $\sigma_1^2 = \mathrm{V}[Y]$ og korrelasjon $\rho = \mathrm{Cov}(X, Y)/(\sigma_1\sigma_2)$. La $Z_1 = (X - \mu_1)/\sigma_1$ og $Z_2 = (Y - \mu_2)/\sigma_2$ være standardiseringene av X og$Y$.

(a) Vis at korrelasjonen mellom Z_1 og Z_2 er lik ρ og identifiser den simultane tettheten til (Z_1, Z_2) som

$$f(z_1, z_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp(-\frac{1}{2(1-\rho^2)}(z_1^2 + z_2^2 - 2\rho z_1 z_2))$$

Hint: Bruk formelen for tettheten til bivariat normalfordeling Devore & Berk, side 258.

(b) Utled den betingede tettheten for Z_2 gitt $Z_1 = z_1$ og påvis at dette er tettheten i $\mathrm{N}(\rho z_1, 1-\rho^2)$-fordelingen.

(c) Vis på denne bakgrunn at

$$\mathrm{E}[Y|X=x] = \mu_2 + \rho\tfrac{\sigma_2}{\sigma_1}(x-\mu_1)$$
$$\mathrm{V}[Y|X=x] = (1-\rho^2)\sigma_2^2$$

Oppgave 5.2. I denne oppgaven skal vi se på hvordan toutvalgsdata kan oppfattes som lineŕegresjon. Anta at forklaringsvariablene x_i er binære, dvs. kan ta verdiene 0 og 1. La videre

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

der $\varepsilon_i \sim \mathrm{N}(0,\sigma^2), i = 1,\ldots,n$. Altså er $\mathrm{E}[Y_i] = \beta_0$ hvis $x_i = 0$ og $\mathrm{E}[Y_i] = \beta_0 + \beta_1$ hvis $x_i = 1$.

(a) Vis at minste kvadraters estimatorene blir $\hat\beta_0 = \bar Y_0$ og $\hat\beta_1 = \bar Y_1 - \bar Y_0$ der $\bar Y_0$ er gjennomsnittet av Y_i-ene med $x_i = 0$ og $\bar Y_1$ er gjennomsnittet av Y_i-ene med $x_i = 1$.

(b) Forklar hvorfor $s^2 = SSE/(n-2) = \sum_{i=1}^n (Y_i - \hat\beta_0 - \hat\beta_1 x_i)^2/(n-2)$ er forventningsrett for σ^2 og kan uttrykkes på samme form som s_p^2 fra Kapittel 9 i Devore & Berk.

(c) Vis at $\mathrm{V}(\hat\beta_1) = \sigma^2(1/n_0 + 1/n_1)$ der n_j er antall observasjoner med $x_i = j, j = 0, 1$.

(d) Forklar hvorfor dette leder til at

$$t_1 = \frac{\hat\beta_1}{s\sqrt{1/n_0 + 1/n_1}} \sim t_{n-2}\text{fordelt under } \mathrm{H}_0 : \beta_1 = 0$$

(e) Benytt dataene om fødselsvekt `fodsler.txt` på kursets hjemmeside (under `data`) og gjør en t-test for forskjell i fødselsvekt mellom gutter og jenter - der du antar lik varians for begge kjønn.

Tilpass deretter en enkel lineær regresjon med Y_i = fødselsvekt og x_i = indikator for kjønn (1=gutter, 0=jenter) og sammenlign med t-testen.

(f) Trekk et subsample på $n = 200$ barn fra det fulle datasettet og gjenta analysen på dette subsamplet.

Oppgave 5.3. Vi ønsker å vise at i en enkel lineær regresjonsmodell $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1,\ldots,n$ der ε_i-ene er uavhengige, har forventninger lik 0 og varianser lik σ^2 så er, med minste kvadraters estimatorene $(\hat\beta_0, \hat\beta_1)$ for (β_0, β_1),

$$s^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^n (Y_i - \hat\beta_0 - \hat\beta_1 x_i)^2}{n-2}$$

forventningsrett for σ^2.

(a) Forklar hvorfor det ikke forandrer på modellen om vi bytter ut x_i-ene med sentrerte forklaringsvariable $x_i' = x_i - \bar x$ der $\bar x$ er gjennomsnittet av x_i-ene.

(b) Angi minste kvadraters estimatorene for β_0 og β_1 når vi har sentrerte forklaringsvariable x_i, dvs. $\sum_{i=1}^n x_i = 0$.

(c) Vis at $SS0 = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2$ har forventning $n\sigma^2$

(d) Vis at med sentrerte forklaringsvariable kan vi uttrykke

$$SSE = SS0 - n(\bar{Y} - \beta_0)^2 - (\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^{n} x_i^2$$

og forklar hvorfor $n(\bar{Y} - \beta_0)^2$ og $(\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^n x_i^2$ begge har forventning σ^2.

Konkluder fra dette om forventningsretthet av s^2.

(e) Anta at $\varepsilon_i \sim \mathrm{N}(0, \sigma^2), i = 1, \ldots, n$ og uavhengige. Vis at $SS0/\sigma^2 \sim \chi^2_n$.

(f) Argumenter for at under forutsetningene i forrige punkt så er $Z_0^2 = n(\bar{Y} - \beta_0)^2/\sigma^2$ og $Z_1^2 = (\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^n x_i^2/\sigma^2$ begge χ^2_1 og uavhengige.

(g) Det kan vises at under forutsetningene i punkt (e) så er $\bar{Y}$ og $\hat{\beta}_1$ uavhenige av $SS0$. Vis på denne bakgrunn at at $SSE/\sigma^2 \sim \chi^2_{n-2}$.

Oppgave 5.4. La $X_1, X_2, \ldots, X_n$ og $Y_1, Y_2, \ldots, Y_m$ være vilkårlige stokastiske variable (med eksisterende varianser).

(a) Vis at

$$\mathrm{Cov}(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{Cov}(X_i, Y_j)$$

(b) La $a, b_1, \ldots, b_n, c$ og $d_1, \ldots, d_m$ være vilkårlige konstanter. Verifiser formel **6.** e) i Formelsamlingen for STK1100 og STK1110

$$\mathrm{Cov}(a + \sum_{i=1}^{n} b_i X_i, c + \sum_{j=1}^{m} d_j Y_j) = \sum_{i=1}^{n}\sum_{j=1}^{m} b_i d_j \mathrm{Cov}(X_i, Y_j)$$

Oppgave 5.5 (Multippel lineær regresjon og matriser)**.** Denne oppgaven er delvis en repetisjon fra gjennomgang på forelesning, men er tatt med siden det ikke direkte er dekket i læreboken. Vi vil gjøre bruk av følgende definisjoner (som også finnes i læreboken):

La $U_1, ..., U_p$ være et sett av tilfeldige variable. Vi sier da at $\mathbf{U} = (U_1, ..., U_p)^T$ er en *tilfeldig vektor*. Videre definerer vi forventingen til $\mathbf{U}$ til å være

$$E(\mathbf{U}) = \begin{pmatrix} E(U_1) \\ E(U_2) \\ \vdots \\ E(U_p) \end{pmatrix}$$

Videre definerer vi *kovariansmatrisen* til $\mathbf{U}$ til å være

$$\mathrm{Cov}(\mathbf{U}) = \begin{pmatrix} V(U_1) & \mathrm{Cov}(U_1, U_2) & \mathrm{Cov}(U_1, U_3) & \cdots & \mathrm{Cov}(U_1, U_p) \\ \mathrm{Cov}(U_2, U_1) & V(U_2) & \mathrm{Cov}(U_2, U_3) & \cdots & \mathrm{Cov}(U_2, U_p) \\ \vdots & & & & \\ \mathrm{Cov}(U_p, U_1) & \mathrm{Cov}(U_p, U_2) & \mathrm{Cov}(U_p, U_3) & \cdots & V(U_p) \end{pmatrix}$$

(a) Vis at for en vilkårlig matrise $\mathbf{A}$ av dimension $q \times p$ og en vektor $\mathbf{b}$ av lengde p så er

$$E(\mathbf{AU} + \mathbf{b}) = \mathbf{A}\mathbf{E}(\mathbf{U}) + \mathbf{b}$$

(b) Vis at

$$\text{Cov}(\mathbf{U}) = \mathbf{E}[(\mathbf{U} - \mathbf{E}(\mathbf{U}))(\mathbf{U} - \mathbf{E}(\mathbf{U}))^{\mathbf{T}}]$$

der vi nå definerer forventningen til en matrise som matrisen av forventningene.

(c) Vis at for en vilkårlig matrise $\mathbf{A}$ av dimension $q \times p$ så er

$$\text{Cov}(\mathbf{AU}) = \mathbf{A}\text{Cov}(\mathbf{U})\mathbf{A}^T$$

Hint: Bruk de to foregående resultatene.

(d) La $\mathbf{A}$ og $\mathbf{B}$ være to matriser med dimensjon $q \times p$ og $r \times p$ henholdsvis. Definer nå $\text{Cov}(\mathbf{AU}, \mathbf{BU})$ til å være matrisen der element (j, k) er kovariansen mellom j'te element av $\mathbf{AU}$ og k'te element av $\mathbf{BU}$. Vis at

$$\text{Cov}(\mathbf{AU}, \mathbf{BU}) = \mathbf{A}\text{Cov}(\mathbf{U})\mathbf{B}^T$$

Hint: Definer en lang tilfeldig vektor som inneholder både $\mathbf{AU}$ og $\mathbf{BU}$.

Oppgave 5.6. Vi vil nå se på multippel lineær regresjon, dvs

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

der $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)$ og $\varepsilon_i \overset{uif}{\sim} N(0, \sigma^2)$. Vi har også at

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

er minste kvadraters estimatoren for β.

(a) Vis at $\widehat{\beta}$ er en forventningsrett estimator for β.

(b) Vis at kovariansmatrisen til $\widehat{\beta}$ er $\sigma^2\mathbf{C}$ der $\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}$.

(c) Argumenter for at $\widehat{\beta}$ er en vektor av normalfordelte variable (og dermed multidimensjonalt normalfordelt).

(d) Hvis $\widehat{\mathbf{Y}}$ er vektoren bestående av $\widehat{Y}_i, i = 1, ..., n$, vis at $\widehat{\mathbf{Y}} = \mathbf{HY}$ der $\mathbf{H}$ er en matrise som er symmetrisk (dvs $\mathbf{H} = \mathbf{H}^T$) og med egenskapene

$$\mathbf{HH} = \mathbf{H};$$
$$[\mathbf{I} - \mathbf{H}][\mathbf{I} - \mathbf{H}] = \mathbf{I} - \mathbf{H}.$$

Her er $\mathbf{I}$ identitetsmatrisen av passende dimensjon.

(e) Vis at $\text{E}(\mathbf{Y} - \widehat{\mathbf{Y}}) = \mathbf{0}$.

(f) Vis at $\text{Cov}(\widehat{\mathbf{Y}}) = \sigma^2\mathbf{H}$ og at $\text{Cov}(\mathbf{Y} - \widehat{\mathbf{Y}}) = \sigma^2[\mathbf{I} - \mathbf{H}]$.

Hva slags fordeling har disse vektorene?

(g) Vis at $\widehat{\beta}$ og $\mathbf{Y} - \widehat{\mathbf{Y}}$ er uavhengige.

Hint: Skriv $\widehat{\beta} = \mathbf{AY}$ og $\mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{BY}$ for passende valg av $\mathbf{A}$ og $\mathbf{B}$ og bruk punkt (d) samt at for normalfordelte variable er kovarians lik null ekvivalent med uavhengighet.

Oppgave 5.7. Denne oppgaven er en utvidelse av denne forrige og viser ytterigere hvordan matriseregning er nyttig i forbindelse med multippel regresjon. Men problemstillingene er kanskje for de mer spesielt interessert og viser hvorfor

$$\hat{\sigma}^2 = \frac{1}{n - (k+1)}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n - (k+1)}(\mathbf{Y} - \widehat{\mathbf{Y}})^T(\mathbf{Y} - \widehat{\mathbf{Y}})$$

er en forventningsrett estimator for σ^2. Vi vil her få bruk for begrepet *trase*, der trasen til en $p \times p$ matrise $\mathbf{V}$, $\text{tr}(\mathbf{V})$ er summen av diagonal-elementene. Vi trenger også at hvis $\mathbf{A}$ er en $p \times q$ matrise og $\mathbf{B}$ er en $q \times p$ matrise, så er

$$\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA}).$$

(a) Vis at hvis $\mathbf{V}$ er en $p \times p$ matrise av tilfeldige variable, så er

$$\text{E}(\text{tr}(\mathbf{V})) = \text{tr}(\text{E}(\mathbf{V}))$$

(b) Vis at $(\mathbf{Y} - \widehat{\mathbf{Y}})^T(\mathbf{Y} - \widehat{\mathbf{Y}}) = \text{tr}((\mathbf{Y} - \widehat{\mathbf{Y}})(\mathbf{Y} - \widehat{\mathbf{Y}})^T)$.

(c) Vis at $E(\text{tr}((\mathbf{Y} - \widehat{\mathbf{Y}})(\mathbf{Y} - \widehat{\mathbf{Y}})^T)) = \sigma^2\text{tr}(\mathbf{I} - \mathbf{H})$.

Hint: Bruk punktene (i) og (j).

(d) Vis at $\text{tr}(\mathbf{H}) = k + 1$ og dermed at $\text{tr}(\mathbf{I} - \mathbf{H}) = n - (k + 1)$.

Bruk dette så til å vise at $E[\hat{\sigma}^2] = \sigma^2$.

Oppgave 5.8. Vi vil i denne oppgaven se nærmere på den multiple lineære regresjonsmodellen:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i; \quad i = 1, 2, ..., n \qquad (*)$$

der x_{ij}-ene er gitte forklaringsvariabler og ε_i-ene er uavhengige og $N(0, \sigma^2)$-fordelte. Det er velkjent at (*) kan skrives på formen $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, der $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_n]^T$, $\varepsilon = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n]^T$, $\beta = [\beta_0, \beta_1, \ldots, \beta_k]^T$ og $\mathbf{X}$ er $n \times (k + 1)$ matrisen der i-te rad har elementene $1, x_{i1}, \ldots, x_{ik}$. Det er også velkjent at minste kvadraters estimator for β er gitt ved $[\mathbf{X}^T\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{Y}$.

(a) Vis at $\hat{\beta}$ er forventningsrett og at kovariansmatrisen til $\hat{\beta}$ er gitt ved $\text{Cov}(\hat{\beta}) = \sigma^2\mathbf{C}$, der $\mathbf{C} = [\mathbf{X}^T\mathbf{X}]^{-1}$.

La $x_1^*, x_2^*, \ldots, x_k^*$ være gitte verdier av forklaringsvariablene. Vi er interessert i å estimere forventet respons svarende til disse verdiene, dvs.

$$\mu(x_1^*, x_2^*, \ldots, x_k^*) = \beta_0 + \beta_1 x_1^* + \beta_1 x_2^* + \cdots + \beta_k x_k^* \qquad (**)$$

Merk at hvis vi innfører vektoren $\mathbf{x}^* = [1, x_1^*, x_2^*, \ldots, x_k^*]^T$, kan (**) gis på formen $\mu(\mathbf{x}^*) = (\mathbf{x}^*)^T\beta$.

(b) Angi en forventningsrett estimator $\hat{\mu}(\mathbf{x}^*)$ for (**) og vis at variansen til estimatoren kan skrives som $V(\hat{\mu}(\mathbf{x}^*)) = \sigma^2 \cdot (\mathbf{x}^*)^T\mathbf{C}\mathbf{x}^*$, der $\mathbf{C}$ er gitt i punkt (a).

(c) Bestem fordelingen til

$$\frac{\hat{\mu}(\mathbf{x}^*) - \mu(\mathbf{x}^*)}{S\sqrt{(\mathbf{x}^*)^T\mathbf{C}\mathbf{x}^*}}$$

der S^2 er den vanlige forventningsrette estimatoren for σ^2. Utled et $100(1 - \alpha)\%$ konfidensintervall for $\mu(\mathbf{x}^*)$.

Hint: Du kan her bruke resultater som er gitt i formelsamlingen.

La Y^* være en ny observasjon fra den lineære regresjonsmodellen (*) svarende til verdiene $\mathbf{x}^* = [x_1^*, x_2^*, \ldots, x_k^*]^T$ av forklaringsvariablene. Vi har altså at

$$Y^* = (\mathbf{x}^*)^T\beta + \varepsilon^*,$$

der ε^* er $\mathrm{N}(0, \sigma^2)$-fordelt og uavhengig av $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$.

(d) Bestem et $100(1-\alpha)\%$ prediksjonsintervall for Y^*.

Oppgave 5.9. Denne oppgaven tar opp et par resultater som brukes mye i forbindelse med matriseformulering knyttet til minste kvadraters estimering.

(a) La $\mathbf{A}$ være en $pxq-$matrise og $\mathbf{B}$ en qxr-matrise. Vis at $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$.

Hint: Skriv ut begge uttrykk på komponentform.

(b) La $\mathbf{A}$ være en pxp-matrise som er inverterbar, dvs. $\mathbf{A}^{-1}$ eksisterer og $\mathbf{A}^T = \mathbf{A}$. Vis at da er også $\mathbf{A}^{-1}$ også symmetrisk.

Hint: Bruk punkt (a) sammen med at $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$, altså en identitetsmatrise.

Oppgave 5.10. (Implementering av multippel lineær regresjon)

Det er tenkelig at man befinner seg på en øde øy med PC uten `R` eller annet egnet program for lineær regresjon, men med et matriseorientert program implementert. Hvis det samtidig oppstår et ustoppelig behov for å tilpasse en multippel lineær regresjonsmodell kan denne oppgaven være til hjelp. Kanskje det også mer generelt er tilfredsstillende at man nokså lett kan programmere estimering, testing og modellsjekking i regresjonsmodeller.

(a) Vi skal bruke dataene om fødselsvekt benyttet på forelesningnene som illustrasjon. Les inn disse og tilpass regresjonsmodellen

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i$$

der Y_i er fødselsvekt, x_{i1} svangerskapslengde, x_{i2} en indikatorvariabel forkjønn, x_{i3} mors alder, x_{i4} antall tidligere fødsler samt ε_i uavhengige feilledd med forventning 0 og varians σ^2.

Tilpass denne modellen i `R` ved `lm`-kommandoen og skriv ut et summary fra denne modellen.

(b) Sett nå opp vektoren av responser $\mathbf{Y}$ og designmatrisen $\mathbf{X}$ for modellen i (a). Beregn minste kvadratersestimatoren $\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ og sammenlign denne med estimatene i (a).

Hint: Du får bruk for `R`-funksjonene `cbind()`, `t()`, `solve()` samt matriseprodukt `%*%`.

(c) Beregn de predikerte verdiene $\hat{\mathbf{Y}} = \mathbf{X}\beta$ samt den residuale kvadratsummen $\mathrm{SSE} = (\mathbf{Y} - Y)^T(\mathbf{Y} - Y)$. Estimer deretter variansen σ^2 med $s^2 = \mathrm{SSE}/(n-4-1)$ og sammenlign med estimatet for σ is (a).

(d) Beregn den estimerte kovariansestimatoren $(\mathbf{X}^T\mathbf{X})^{-1}s^2$ og finn diagonalelementene i denne med `diag()`. Sammenlign med standardfeilene fra (a).

(e) Finn total kvadratsum SST og beregn multippel R^2 samt F-observatoren for om noen av $\beta_j \neq 0, j = 1, 2, 3, 4$.

(f) Beregn deretter hatt-matrisen $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ og finn dennes dimensjon med `dim()`. Lag en sammenfatning av influens- eller leverageverdiene h_{ii} langs diagonalen til $\mathbf{H}$.

(g) Beregn tilslutt de standardiserte residualene $e_i^* = (Y_i - \hat{Y}_i)/(s\sqrt{1-h_{ii}})$ og generer de fire vanlige residualplottene i `R`.

(Det er ikke så nøye med aksetekster, tilpassede linjer, etc., men sjekk at det er tilsvarende plott som ved `R`-kommandoen `plot()` andvendt på et `lm`-objekt.)

Oppgave 5.11. New York by er full av italienske restauranter. De varierer i pris og kvalitet. Vi skal studere hvordan pris betalt for en middag på italiensk restaurant i NYC kan forklares av hvor god maten er, kombinert med grad av service og eleganse (dekor). Alle disse tre forklaringsvariablene er fastsatt på en kontinuerlig skala av restaurantkritikere, der lav score er dårlig og høy score er bra. Vi har data fra $n = 168$ restaurantbesøk, og tilpasser en multippel regresjonsmodell i R med følgende resultat:

```
Call:
lm(Price ~ Food + Service + Decor, data = nyc)

Residuals:
   Min     1Q Median     3Q    Max
-14.8440 -3.7039 -0.1525 3.6218 19.0576

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.6409     4.7536  -5.184 6.33e-07
Food          1.5556     0.3731   4.170 4.93e-05
Service       0.1350     0.3957   0.341    0.733
Decor         1.8473     0.2176   8.491 1.17e-14
---

Residual standard error: 5.803 on 164 degrees of freedom
Multiple R-squared: 0.617,Adjusted R-squared: 0.61
F-statistic: 88.06 on 3 and 164 DF,  p-value: < 2.2e-16
```

Her betegner ???Food??? score for kvaliteten på maten.

(a) Sett opp modellen som ligger til grunn for analysen. Bruk matriseform. Benytt notasjonen $\mathbf{Y}$ for responsvektor, $\mathbf{X}$ for designmatrisen og β for parametervektoren. Angi dimensjonen til alle vektorer/matriser. Formuler de vanlige forutsetningene vi gjør for en slik modell. Forklar kort hva de estimerte koeffisientene sier om sammenhengen mellom responsen og de tilhørende forklaringsvariablene.

(b) Vis at minste kvadraters estimator $\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ er forventningsrett for β. Begrunn kort hvorfor $\hat{\beta}_j$-ene blir normalfordelte. Utled et uttrykk for kovariansmatrisen til $\widehat{\beta}$, $\mathrm{Cov}(\widehat{\beta})$. Du kan bruke at dersom $\mathbf{A}$ er en matrise med konstanter og $\mathbf{V} = \mathbf{AU}$, så er $\mathrm{Cov}(\mathbf{V}) = \mathbf{A}\mathrm{Cov}(\mathbf{U})\mathbf{A}$.

(c) Utfør, trinn for trinn, en hypotesetest for å teste om kvaliteten p?? maten (Food), under ellers like omstendigheter (service og dekor), har en signifikant positiv effekt på prisen. Du kan bruke tall fra R-utskriften. Skriv en konklusjon.

(d) Multippel R^2 er definert ved

$$R^2 = 1 - \frac{SSE}{SST}$$

der $SSE = \sum_i (Y_i - \widehat{Y}_i)^2$ og $SST = \sum_i (Y_i - \bar{Y}^2$. Forklar hvordan denne skal tolkes. Justert R^2 , R^2_{adj}, er definert som

$$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

der k er antall kovariater i modellen. Diskuter hvorfor denne kan være å foretrekke fremfor R^2 i mange (hvilke?) situasjoner.

(e) Hvilke hypoteser testes i siste linje i R-utskriften (F-test), og hvorledes blir konklusjonen?