

Eksamens i Statistisk Dataanalyse 1 (LMAB10069)
November 2012

Alle sædvanlige hjælpemidler er tilladt, herunder bøger og lommeregner men ikke PC. Der er 6 sider med 3 opgaver med i alt 11 spørgsmål, der alle ønskes besvaret. De tre opgaver indgår med samme vægt i bedømmelsen. Benyttede statistiske modeller skal opskrives, og hypoteser, der testes, skal specificeres både udtrykt ved modellens parametre og i ord. Der er i opgaverne givet beregningshjælp i form af (dele af) udskrifter fra R. Dele heraf, men ikke nødvendigvis det hele, skal bruges.

Opgave 1.

Som et led i en undersøgelse af mekanismer bag resistens overfor en svamp (*Septoria tritici*) undersøges forekomst af H_2O_2 i et antal celler fra hver af to plantesorter: Sevin, som er modtagelig over for svampen, og Stakado, som er resistent. Nedenstående tabel viser for hver af de to sorter antallene af celler henholdsvis med og uden forekomst af H_2O_2 .

Stakado		Sevin	
Antal celler		Antal celler	
med H_2O_2	uden H_2O_2	med H_2O_2	uden H_2O_2
49	1402	27	1408

- Beregn for hver af de to sorter et 95% konfidensinterval for brøkdelen af celler med H_2O_2 .
- Man er interesseret i om der er forskel på forekomsten af H_2O_2 fra den ene art til den anden. Opskriv en hypoteze, som kan benyttes til at undersøge dette spørgsmål. Beregn en teststørrelse for test af hypotesen, og angiv dens fordeling. P-værdien ønskes ikke beregnet.
- Uanset ovenstående data regnes nu med, at p-værdien for det udførte test bliver 0.012. Hvad konkluderer du? (Her ønskes blot en eller to sætninger.)
- Angiv et estimat og et konfidensinterval for odds-ratio for forekomst af H_2O_2 på Stakado i forhold til Sevin.

Opgave 2.

Koncentrationen af frie fedtsyrer i blodet måltes hos 32 personer umiddelbart før et måltid (i det følgende kaldt fastemålingen) og igen 4 timer efter måltidet. I R-programmet sidst i opgaven er disse målinger betegnet henholdsvis y_0 og y_4 . Modelkontrol ønskes ikke foretaget i denne opgave.

- Kan der på det foreliggende grundlag påvises nogen sammenhæng mellem fastemålingen og målingen efter måltidet? Angiv ved besvarelse af dette spørgsmål den statistiske model, du benytter, eventuel(le) hypotese(r) og formulér en konklusion.

- (b) Kan der påvises nogen ændring i koncentrationen af frie fedtsyrer fra fastemålingen til målingen 4 timer efter måltidet? Angiv også her den statistiske model, du benytter, eventuel(le) hypotese(r) og formulér en konklusion.
- (c) Angiv et interval som vil indeholde målingen 4 timer efter måltidet for 95% af personerne i populationen i den tænkte situation, at alle personer i populationen måles som beskrevet ovenfor.
- (d) Angiv tilsvarende et interval som vil indeholde målingen 4 timer efter måltidet for 95% af de personer i populationen, hvis fastemåling er 6.0. Som svar her skal du angive en formel for intervallet og desuden indsætte de værdier, du kender, i formlen, men udregning af intervallet kræves ikke.

R-udskriften nedenfor indledes for fuldstændighedens skyld med udskrift af vektorerne y_0 og y_4 . Det er ikke nødvendigt at benytte tallene heri for at besvare opgaven.

```
> y0
[1] 5.87 6.44 6.11 6.43 6.36 6.71 5.85 6.27 6.49 6.32 6.60 6.78 6.60 6.41 6.56
[16] 6.80 5.55 6.64 6.43 6.28 5.74 6.47 5.64 5.56 6.48 6.30 6.16 6.67 5.83 6.33
[31] 6.14 6.28
> y4
[1] 5.81 6.62 6.08 5.28 5.88 6.34 5.08 5.58 5.85 6.24 6.77 5.58 6.90 5.67 6.63
[16] 6.20 5.49 6.15 4.86 5.53 5.61 5.14 5.38 4.48 5.74 5.82 5.20 5.02 4.96 6.35
[31] 5.52 6.28

> mean(y0)
[1] 6.284375
> mean(y4)
[1] 5.75125
> sd(y0)
[1] 0.3528313
> sd(y4)
[1] 0.5936804

> m1 <- lm(y4 ~ y0)
> summary(m1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.5870     1.6872    0.348   0.73034
y0          0.8218     0.2681    3.065   0.00457 **
---
Residual standard error: 0.5266 on 30 degrees of freedom
Multiple R-squared:  0.2385,    Adjusted R-squared:  0.2131
F-statistic: 9.397 on 1 and 30 DF,  p-value: 0.004569
```

```

> m2 <- lm(y4 ~ y0 - 1)
> summary(m2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
y0    0.91488     0.01458   62.75   <2e-16 ***
---
Residual standard error: 0.5191 on 31 degrees of freedom
Multiple R-squared:  0.9922,    Adjusted R-squared:  0.9919
F-statistic: 3937 on 1 and 31 DF,  p-value: < 2.2e-16

> t.test(y0,y4, paired= FALSE, var.equal=TRUE)

Two Sample t-test

data: y0 and y4
t = 4.3669, df = 62, p-value = 4.875e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
0.289082 0.777168
sample estimates:
mean of x mean of y
6.284375 5.751250

> t.test(y0,y4, paired= FALSE, var.equal=FALSE)

Welch Two Sample t-test

data: y0 and y4
t = 4.3669, df = 50.47, p-value = 6.264e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
0.287968 0.778282
sample estimates:
mean of x mean of y
6.284375 5.751250

```

```

> t.test(y0,y4, paired= TRUE)

  Paired t-test

data: y0 and y4
t = 5.7789, df = 31, p-value = 2.305e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
0.344972 0.721278
sample estimates:
mean of the differences
0.533125

> qnorm(0.975)
[1] 1.959964
> qt(0.975, df=62)
[1] 1.998972
> qt(0.975,df=31)
[1] 2.039513
> qt(0.975,df=30)
[1] 2.042272
> qt(0.975,df=15)
[1] 2.13145

```

Opgave 3.

Af de 32 forsøgspersoner fra opgave 2 var de første 16 personer ældre og de sidste 16 personer yngre, og i hver af de to aldersgrupper var de 8 første mænd, mens de 8 sidste var kvinder. I udskriften fra R sidst i opgaven betegner kon forsøgspersonens køn (enten Male eller Female), mens alder betegner aldersgruppen, idet A står for yngre, og B står for ældre. Vektorerne y0 og y4 er som i opgave 2, men løsning af opgave 2 er ikke nødvendig for at løse denne opgave. Ligesom i opgave 2 ønskes modelkontrol ikke foretaget i denne opgave, og det er ikke nødvendigt at benytte tallene i vektorerne y0 og y4 for at besvare opgaven.

- (a) Opskriv en statistisk model som udgangspunkt for at undersøge, om koncentrationen (y4) af frie fedtsyrer 4 timer efter måltidet afhænger af alder og køn. Angiv estimatet for spredningen i modellen.
- (b) Undersøg ved hjælp af R-udskriften sidst i opgaven spørgsmålet om koncentrationen (y4) af frie fedtsyrer 4 timer efter måltidet afhænger af køn og alder. For det eller de test, du bruger i besvarelserne, skal hypotesen opskrives, og der ønskes præcis angivelse af teststørrelse og P-værdi.
- (c) Formulér en kort og klar konklusion og angiv estimat(er) og konfidensinterval(ler) som du finder passende på baggrund af din konklusion.

```

> model1 <- lm(y4 ~ 1)
> model2 <- lm(y4 ~ kon)
> model3 <- lm(y4 ~ alder)
> model4 <- lm(y4 ~ kon - 1)
> model5 <- lm(y4 ~ alder - 1)
> model6 <- lm(y4 ~ kon + alder)
> model7 <- lm(y4 ~ kon*alder)

> summary(model1)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.7512     0.1049   54.8   <2e-16 ***
---
Residual standard error: 0.5937 on 31 degrees of freedom

> summary(model2)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.9206     0.1444   41.003  <2e-16 ***
konMale     -0.3387    0.2042   -1.659    0.108
---
Residual standard error: 0.5776 on 30 degrees of freedom
Multiple R-squared: 0.08402,      Adjusted R-squared: 0.05349
F-statistic: 2.752 on 1 and 30 DF,  p-value: 0.1076

> summary(model3)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.4706     0.1323   41.339  <2e-16 ***
alderB      0.5612     0.1872    2.999   0.0054 **
---
Residual standard error: 0.5293 on 30 degrees of freedom
Multiple R-squared: 0.2306,      Adjusted R-squared: 0.205
F-statistic: 8.993 on 1 and 30 DF,  p-value: 0.005405

> summary(model4)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
konFemale   5.9206     0.1444   41.00   <2e-16 ***
konMale     5.5819     0.1444   38.66   <2e-16 ***
---
Residual standard error: 0.5776 on 30 degrees of freedom
Multiple R-squared: 0.9906,      Adjusted R-squared: 0.99
F-statistic: 1588 on 2 and 30 DF,  p-value: < 2.2e-16

```

```

> summary(model5)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
alderA     5.4706    0.1323   41.34 <2e-16 ***
alderB     6.0319    0.1323   45.58 <2e-16 ***
---
Residual standard error: 0.5293 on 30 degrees of freedom
Multiple R-squared: 0.9921,      Adjusted R-squared: 0.9916
F-statistic: 1893 on 2 and 30 DF, p-value: < 2.2e-16

> summary(model6)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.6400    0.1556   36.250 < 2e-16 ***
konMale     -0.3388    0.1797  -1.886  0.06941 .
alderB      0.5612    0.1797   3.124  0.00403 **
---
Residual standard error: 0.5081 on 29 degrees of freedom
Multiple R-squared: 0.3147,      Adjusted R-squared: 0.2674
F-statistic: 6.657 on 2 and 29 DF, p-value: 0.004175

> summary(model7)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.6113    0.1825   30.744 <2e-16 ***
konMale     -0.2812    0.2581  -1.090  0.2852
alderB      0.6188    0.2581   2.397  0.0234 *
konMale:alderB -0.1150    0.3650  -0.315  0.7551
---
Residual standard error: 0.5162 on 28 degrees of freedom
Multiple R-squared: 0.3171,      Adjusted R-squared: 0.2439
F-statistic: 4.333 on 3 and 28 DF, p-value: 0.0125

> qt(0.975,df=31)
[1] 2.039513
> qt(0.975,df=30)
[1] 2.042272
> qt(0.975,df=29)
[1] 2.04523
> qt(0.975,df=28)
[1] 2.048407

```

Eksamensopgave i Statistisk Dataanalyse 1 (LMAB10069)

Januar 2013

Alle sædvanlige hjælpemidler er tilladt, herunder bøger og lommeregner men ikke PC. Der er 6 sider med 3 opgaver med i alt 15 spørgsmål, der alle ønskes besvaret. Den første opgave indgår med vægten 40 procent i bedømmelsen, de to andre opgaver med hver 30 procent. Benyttede statistiske modeller skal opskrives, og hypoteser, der testes, skal specificeres både udtrykt ved modellens parametre og i ord. Der er sidst i hver opgave givet beregningshjælp i form af (dele af) udskrifter fra R. Dele heraf, men ikke nødvendigvis det hele, skal bruges.

Opgave 1. For at undersøge effektiviteten af sprøjting af planter målte man i en eksperimentel opstilling den hastighed, hvormed draberne ramte planternes blade under forskellige omstændigheder. I denne opgave benyttes et datasæt med 46 sammenhørende målinger af dråbehastigheden (v) og dråbediameteren (d). Ifølge en teoretisk model (baseret på Stokes lov), skulle logaritmen til dråbehastigheden ($\ln v$ i R-programmet) afhænge lineært af logaritmen til dråbediameteren ($\ln d$ i R-programmet).

- Opskriv modellen svarende til ovennævnte teori, og angiv estimerater for samtlige parametre i modellen.
- Angiv et konfidensinterval for hældningen i samme model.
- Sidst i R-programmet er der to anvendelser af funktionen `anova`, som resulterer i tilsammen tre F-test, som hver fylder en linie i udskriften. Opskriv de dertil hørende modeller og hypoteser, og skriv en kort konklusion for hvert af de tre test.
- Figuren, som udskrives sidst i programmet, er til brug for modelkontrol. Skriv kort, gerne i punktform, hvilke(n) antagelse(r) denne figur er velegnet til at kontrollere, og hvad du vil konkludere om modellen ud fra figuren.
- Ved en fejltagelse blev det foreslået at køre modellen

```
> fejlmodel <- lm(lnv ~ factor(lnd))
> summary(fejlmodel)
```

Udskriften var imidlertid ikke brugbar og blandt andet kunne residualspredningen (residual standard error) ikke beregnes. Opgiv den benyttede model, og forklar kort, hvorfor residualspredningen i den ikke kan estimeres.

```
> head(hast)
   d    v
1 210  35
2 264 420
3 300 162
```

```

4 396 510
5 252 278
6 240 185

> dim(hast)
[1] 46 2
> attach(hast)
> lnd <- log(d)
> lnv <- log(v)

> lm0 <- lm(lnv ~ lnd - 1)
> summary(lm0)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
lnd    0.81058    0.02361   34.33  <2e-16 ***
Residual standard error: 0.8467 on 45 degrees of freedom
Multiple R-squared: 0.9632,      Adjusted R-squared: 0.9624
F-statistic: 1179 on 1 and 45 DF, p-value: < 2.2e-16

> lm1 <- lm(lnv ~ lnd)
> summary(lm1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.6232     1.5463  -5.577 1.41e-06 ***
lnd         2.4382     0.2924   8.338 1.33e-10 ***
Residual standard error: 0.6554 on 44 degrees of freedom
Multiple R-squared: 0.6124,      Adjusted R-squared: 0.6036
F-statistic: 69.52 on 1 and 44 DF, p-value: 1.327e-10

> lnd2 <- lnd*lnd
> lm2 <- lm(lnv ~ lnd + lnd2)
> summary(lm2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.9437    11.6651   2.653  0.01114 *
lnd        -13.2388     4.5966  -2.880  0.00617 **
lnd2        1.5439     0.4519   3.416  0.00140 **
Residual standard error: 0.588 on 43 degrees of freedom
Multiple R-squared: 0.6951,      Adjusted R-squared: 0.681
F-statistic: 49.02 on 2 and 43 DF, p-value: 8.094e-12

```

```

> lm3 <- lm(lnv ~ lnd2 - 1)
> summary(lm3)

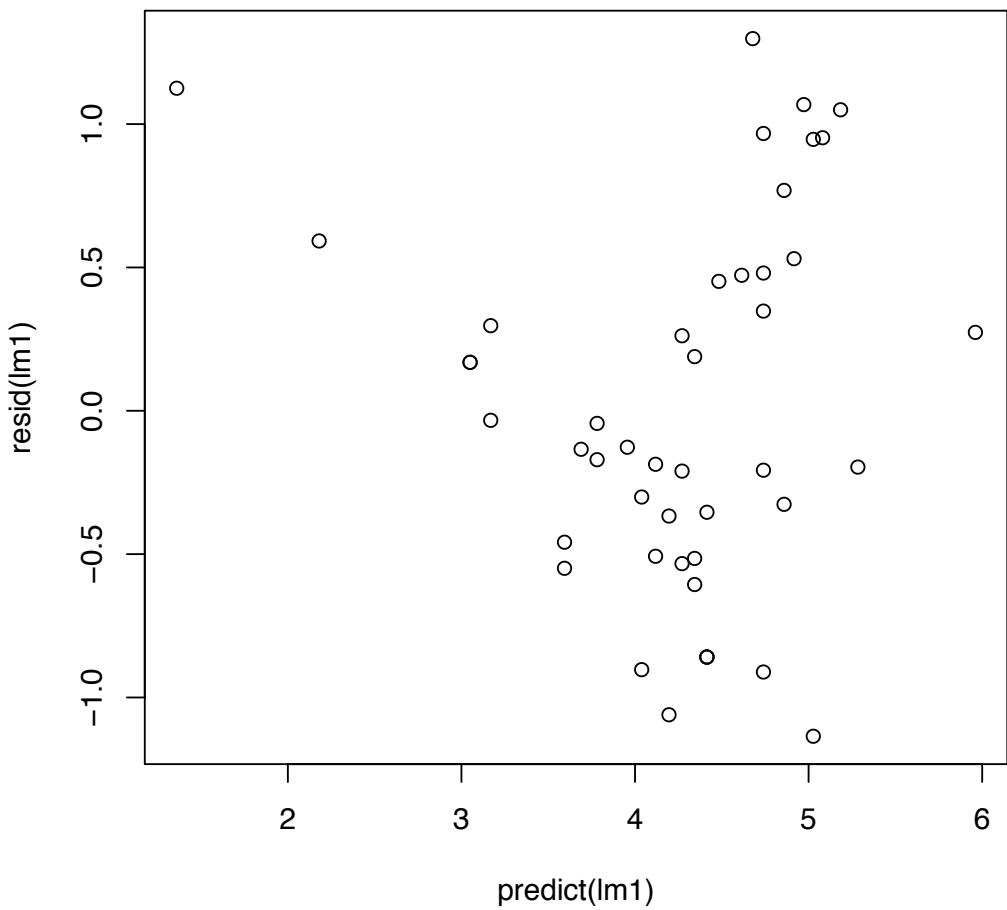
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
lnd2 0.153112   0.003671    41.7   <2e-16 ***
Residual standard error: 0.7012 on 45 degrees of freedom
Multiple R-squared:  0.9748,    Adjusted R-squared:  0.9742
F-statistic: 1739 on 1 and 45 DF,  p-value: < 2.2e-16

> anova(lm0, lm1, lm2)
Analysis of Variance Table
Model 1: lnv ~ lnd - 1
Model 2: lnv ~ lnd
Model 3: lnv ~ lnd + lnd2
  Res.Df   RSS Df Sum of Sq    F    Pr(>F)
1     45 32.263
2     44 18.902  1   13.361 38.643 1.771e-07 ***
3     43 14.867  1    4.035 11.670  0.001398 **

> anova(lm3, lm2)
Analysis of Variance Table
Model 1: lnv ~ lnd2 - 1
Model 2: lnv ~ lnd + lnd2
  Res.Df   RSS Df Sum of Sq    F    Pr(>F)
1     45 22.125
2     43 14.867  2    7.258 10.496 0.000194 ***

> plot(predict(lm1), resid(lm1))

```



Opgave 2.

Hvert af 112 insekter blev smittet med en svampesygdom. Der var to stammer (A og B) af svamphen. Den ene halvdel af insekterne smittedes med den ene stamme og den anden halvdel smittedes med den anden stamme. Efter tre dage talte man, hvor mange af insekterne i hver gruppe, der var døde. Der var 38 (ud af 56) døde for stamme A, mens der var 28 (ud af 56) døde for stamme B.

Disse resultater var en del af et større forsøg, og man var blandt andet interesseret i, om der var grund til at skelne mellem de to stammer i den videre analyse.

- Kan man, på basis af ovenstående resultater, med rimelighed påstå, at der er forskel på de to stammer, hvad angår dødeligheden af insekterne? Hvis svaret baseres på et statistisk test, kræves teststørrelsen (test statistic) beregnet, men p-værdien kræves

ikke beregnet, og du må som grundlag for konklusionen lade som om, at p-værdien er 0.08.

Den videre analyse indeholdt en undersøgelse af insekternes overlevelsestid. Hertil udførtes en logistisk regressionsanalyse på basis af, at der ud af de 112 insekter var henholdsvis 1, 8, 66, 99 og 110 der døde efter henholdsvis 1, 2, 3, 4 og 5 dage.

- b) Benyt denne analyse til at angive et estimat for det tidsrum, som halvdelen af insekterne overlever.
- c) Benyt ligeledes denne analyse til at angive et estimat for odds-ratio for at overleve i fire dage i forhold til at overleve i to dage.
- d) Angiv desuden et estimat for samme odds-ratio uden at benytte den logistiske regressionsmodel.
- e) Den logistiske regressionsanalyse udført på de foreliggende data er særdeles kritisabel, da en eller flere af antagelserne er klart forkert(e). Hvilke(n)?

```
> day <- c(1, 2, 3, 4, 5)
> total <- c(112, 112, 112, 112, 112)
> dead <- c(1, 8, 66, 99, 110)
> alive <- total - dead
> survival <- matrix(c(alive, dead), nrow=5, ncol=2)

> glm1 <- glm(survival ~ day, family=binomial)
> summary(glm1)
```

Call:
`glm(formula = survival ~ day, family = binomial)`

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	6.6038	0.5611	11.77	<2e-16 ***
day	-2.2278	0.1831	-12.17	<2e-16 ***

Null deviance: 454.9820 on 4 degrees of freedom
Residual deviance: 5.0422 on 3 degrees of freedom

Opgave 3.

Ved økonomiske investeringer og handler er det normalt, at en højere forventet gevinst er sammenkoblet med en forøget risiko. I denne opgave antager vi at gevinsten ved en investering af type A er normalfordelt med middelværdi (mean) μ_A og spredning (standard deviation) σ_A , mens gevinsten ved en investering af type B er normalfordelt med middelværdi μ_B og spredning σ_B . Bemærk at “gevinsten” i visse tilfælde kan være negativ.

- (a) Antag at $\mu_A = 10000$ og $\sigma_A = 5000$. Angiv et interval hvori gevinsten ved en investering af type A ligger med sandsynlighed 0.95.
- (b) Hvor stor er sandsynligheden for, at gevinsten ved en investering af type A bliver negativ?
- (c) Antag nu endvidere at $\mu_B = 15000$ og $\sigma_B = 9000$. Hvis der foretages en investering af hver type, hvad er da sandsynligheden for, at investeringen af type B giver større gevinst end investeringen af type A?
- (d) Antag nu, at der bliver mulighed for en investering af en ny type, C, som har middelværdi $\mu_C = 20000$. Hvor stor må spredningen, σ_C være uden at risikoen for tab bliver større end ved en investering af type A?
- (e) Ved 10 investeringer af endnu en ny type, D, var gevinsterne

8112, 3899, 4584, 3593, 1390, 2346, 8658, 5831, 10499, -227

Beregn et 95% konfidensinterval for middelværdien af gevinsten ved denne type investering. Husk at opskrive den statistiske model, du bruger hertil.

```
> qnorm(0.05)
[1] -1.644854
> qnorm(0.95)
[1] 1.644854
> qnorm(0.975)
[1] 1.959964
> pnorm(0)
[1] 0.5
> pnorm(-5)
[1] 2.866516e-07
> pnorm(2)
[1] 0.9772499
> pnorm(5000/(15000))
[1] 0.6305587
> pnorm(5000/(14000))
[1] 0.6395076
> pnorm(5000/(4000))
[1] 0.8943502
> pnorm(5000/sqrt(5000**2 + 9000**2))
[1] 0.6863898

> y= c(8112, 3899, 4584, 3593, 1390, 2346, 8658, 5831, 10499, -227)
> mean(y)
[1] 4868.5
> sd(y)
[1] 3408.884
```

Eksamens i Statistisk Dataanalyse 1 (LMAB10069U)
November 2013

Alle hjælpemidler er tilladt, herunder lommeregner, men ikke PC. Der er 4 sider med 3 opgaver, som har henholdsvis 3, 5 og 5 spørgsmål, der alle ønskes besvaret. De tre opgaver indgår med vægtene 30%, 35% og 35% i bedømmelsen. Benyttede statistiske modeller skal opskrives, og hypoteser, der testes, skal specificeres både udtrykt ved modellens parametre og i ord. Der er i opgaverne givet beregningshjælp i form af (dele af) udskrifter fra R. Dele heraf skal benyttes, men der kan også være overflødige dele.

Opgave 1.

Er det usædvanligt at hedde det samme til fornavn og efternavn, som f.eks. *Jens Jensen*? For at besvare dette spørgsmål kan man søge information om danskernes navne på *Danmarks Statistik*'s hjemmeside. Følgende 2×2 tabel viser hvor mange danske mænd, der i 2013 hed en af de fire kombinationer af fornavnene *Jens* og *Hans*, og efternavnene *Jensen* og *Hansen*:

Fornavn	Efternavn	
	<i>Jensen</i>	<i>Hansen</i>
<i>Jens</i>	3752	1997
<i>Hans</i>	2160	2807

- Beregn det forventede antal mænd med hver af de 4 navnekombinationer under antagelse af, at der ikke er nogen sammenhæng mellem for- og efternavn.
- Test nulhypotesen at der ikke er nogen sammenhæng mellem for- og efternavn i tabellen ovenfor. Beskriv konklusionen af testet i ord.
- Angiv et estimat og et 95% konfidensinterval for odds ratio for at hedde *Jens* frem for *Hans*, hvis man hedder *Jensen* i forhold til hvis man hedder *Hansen*.

Opgave 2.

For 20 grise udvalgt tilfældigt fra en population har man med to forskellige metoder (A og B) forsøgt at måle graden af infektion. Resultaterne (to målinger for hver af de 20 grise) fremgår af R-udskriften sidst i opgaven. Benyt relevante dele af udskriften ved besvarelse af opgaven. Modelvalidering ønskes ikke foretaget i denne opgave.

- Kan du på det foreliggende grundlag slutte, at der er forskel på de to målemetoder? Angiv ved besvarelse af dette spørgsmål den statistiske model, du benytter, eventuel(le) hypoteze(r) og formulér en konklusion.
- Angiv et interval som med passende sikkerhed indeholder forskellen mellem de to målingers populationsmiddelværdi. (Vælg selv meningen med "passende sikkerhed".)

- (c) Angiv et interval som med passende sikkerhed vil indeholde forskellen mellem de to typer målinger for en tilfældigt valgt gris i populationen. (Vælg igen selv meningen med "passende sikkerhed".)
- (d) Angiv, med udgangspunkt i modellen fra spørgsmål (a), et estimat for sandsynligheden for, at målemetode A fører til et mindre resultat end målemetode B for en tilfældigt udvalgt gris fra populationen.
- (e) Angiv et estimat og et konfidensinterval for samme sandsynlighed som i spørgsmål (d), men nu uden antagelse om normalfordeling. (Vejledning: her er tale om en bestemt hændelse indtræffer eller ej for den udvalgte gris.)

R-udskriften nedenfor indledes med udskrift af vektorerne A og B, som indeholder resultaterne fra målemetode A henholdsvis B for gris nummer 1, 2, ..., 20 samt af vektoren med differencerne.

```
> A
[1] 28.0 56.5 19.5 27.5 61.0 31.5 53.0 58.0 38.0 38.0 50.5 49.5 53.5 22.5 36.5
[16] 22.5 30.0 42.5 14.0 44.0
> B
[1] 21.3 38.4 17.0 20.4 39.8 33.4 38.5 37.7 21.7 16.0 45.8 37.7 42.5 8.9 27.0
[16] 20.7 26.9 36.8 11.3 41.9
> dif <- A-B
> dif
[1] 6.7 18.1 2.5 7.1 21.2 -1.9 14.5 20.3 16.3 22.0 4.7 11.8 11.0 13.6 9.5
[16] 1.8 3.1 5.7 2.7 2.1

> mean(A)
[1] 38.825
> mean(B)
[1] 29.185
> mean(dif)
[1] 9.64
> sd(A)
[1] 14.11445
> sd(B)
[1] 11.35825
> sd(A-B)
[1] 7.291769

> model1 <- lm(A ~ B)
> summary(model1)
```

Call:

```

lm(formula = A ~ B)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.7128    4.6982   1.642   0.118
B            1.0660    0.1505   7.083 1.33e-06 ***
---
Residual standard error: 7.452 on 18 degrees of freedom
Multiple R-squared: 0.7359,      Adjusted R-squared: 0.7213
F-statistic: 50.16 on 1 and 18 DF,  p-value: 1.327e-06

```

Opgave 3.

Rødkløver har ofte så lange kronrør, at de nektarsamlende biers tunger er for korte til at kunne nå ned gennem kronrøret. For at fremme bestøvningen af rødkløveren har man derfor forsøgt at forkorte kronrørene ved brug af vækstreguleringsmidler. Nærværende data omhandler alene midlet Alar, som blev anvendt i koncentrationerne 0, 1.5, 3.0 eller 6.1 ved en sprøjtning tidligt i vækstsæsonen. Planterne blev dyrket i kar, og der var to kar for hver af koncentrationerne. Som resultat ses under navnet lgd i R-programmet nedenfor gennemsnitslængden af de målte kronrør for hvert af de otte kar, mens konc angiver de tilhørende koncentrationer.

```

> lgd
[1] 9.84 10.07 9.59 9.52 8.93 9.34 8.55 8.67
> konc
[1] 0.0 0.0 1.5 1.5 3.0 3.0 6.1 6.1

```

```

> m1= lm(lgd ~ konc)
> summary(m1)

```

Call:
`lm(formula = lgd ~ konc)`

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	9.89491	0.08737	113.258	3.19e-11
konc	-0.21931	0.02510	-8.737	0.000124

Residual standard error: 0.1602 on 6 degrees of freedom
Multiple R-squared: 0.9271, Adjusted R-squared: 0.915
F-statistic: 76.34 on 1 and 6 DF, p-value: 0.0001244

- (a) Opskriv en statistisk model for kronrørslængderne og benyt ovenstående R-udskrift til at estimere parametrene i modellen. Hvilken af parametrene siger noget om, hvorvidt sprøjtningen påvirker kronrørslængden? Angiv et konfidensinterval for denne parameter.

- (b) Du skal nu tage stilling til hvilke af følgende fem udsagn, der er rimelige på basis af ovenstående analyse. Skriv (uden begrundelse) i din besvarelse en indrammet linie med nummeret/numrene på det/de korrekte udsagn.
1. Analysen viser, pånær tilfældigheder, ikke nogen sammenhæng mellem koncentration og kronrørlængde.
 2. Analysen viser klart, at kronrørlængderne gennemgående bliver længere ved stigende koncentration.
 3. Analysen viser klart, at kronrørlængderne gennemgående bliver kortere ved stigende koncentration.
 4. Analysen viser klar sammenhæng mellem koncentration og kronrørlængde, men man kan ikke af udskriften se i hvilken retning sammenhængen går.
 5. Analysen viser, at sammenhængen mellem koncentration og kronrørlængde følger en ret linie.
- (c) Angiv et estimat for den koncentration, der giver en forventet kronrørlængde på 9 mm.
- (d) Angiv et estimat for den koncentration, der med 95% sandsynlighed giver en kronrørlængde på ikke over 9 mm. [Vejledning: For enhver koncentration kan middelværdi og spredning af kronrørlængden udtrykkes ved parametrene, som du blot erstatter med deres estimerater i den videre beregning.]
- (e) I forsøget blev der i nogle kar også sprøjtet på et sent tidspunkt i vækstsæsonen. Lad `konc1` betegne koncentrationerne ved den første sprøjtning og `konc2` koncentrationerne ved den anden sprøjtning. Skriv en eller nogle få liniers R-kode, som du kunne tænke dig at bruge til analyse af dette udvidede forsøg. I besvarelseren ønskes kun R-koden — ikke forklaring på, hvordan analysen udføres. [Specificér selv eventuelle manglende oplysninger om forsøget.]

Reeksamen i Statistisk Dataanalyse 1 (LMAB10069)

Januar 2014

Alle hjælpemidler er tilladt, herunder lommeregner, men ikke PC. Der er 6 sider med 3 opgaver med i alt 12 spørgsmål, der alle ønskes besvaret. Opgave 1 indgår med vægten 40% i bedømmelsen, mens de to andre opgaver hver indgår med vægten 30%. Benyttede statistiske modeller skal opskrives, og hypoteser, der testes, skal specificeres både udtrykt ved modellens parametre og i ord. Der er i opgaverne givet beregningshjælp i form af (dele af) udskrifter fra R. Dele heraf, men ikke nødvendigvis det hele, skal bruges.

Opgave 1.

I et forsøg med etablering af elefantgræs dyrkedes elefantgræs af alder 1, 2 eller 3 år i kar med en af tre mulige jordtyper (A: humusjord, B: lerjord, C: sandjord). Ved slutningen af vækstsæsonen vejedes tørstofmængden af toppen for hvert kar (alle plantedele mere end 5 cm over jorden). Der var i alt 36 kar, nemlig 4 kar for hver kombination af alder og jordtype, og man er interesseret i at undersøge hvordan tørstofvægten af toppen afhænger af jordtype og alder.

I R-programmet i slutningen af denne opgave angiver vektoren `top` tørstofvægten af toppen (i gram).

- (a) Opskriv en statistisk model som udgangspunkt for at undersøge, hvordan tørstofvægten af toppen afhænger af jordtype og alder, herunder om effekten af alder kan vises at afhænge af jordtypen. Modellen skal svare til en af modellerne fra R-programmet.
- (b) I modellen gøres nogle antagelser. Skriv (uden begrundelse) i din besvarelse en indrammet linie med numrene på de af nedenstående forslag til antagelser, som faktisk er antagelser i modellen. Du skal ikke her tage stilling til, om antagelserne er rimelige for denne model, blot om det er antagelser, som modellen er baseret på.
 1. Middelværdien af tørstofvægten af toppen afhænger lineært af jordtype og alder.
 2. Middelværdien af tørstofvægten af toppen afhænger lineært af alder.
 3. Tørstofvægten af toppen er normalfordelt med en middelværdi, som afhænger af jordtype og alder.
 4. Tørstofvægten af toppen er normalfordelt med en middelværdi, som ikke afhænger af jordtype og alder.
 5. Tørstofvægten af toppen er normalfordelt med en spredning (standard deviation), som afhænger af jordtype og alder.
 6. Tørstofvægten af toppen er normalfordelt med en spredning (standard deviation), som ikke afhænger af jordtype og alder.
 7. Alle variablene tørstofvægt af top, jordtype og alder er normalfordelt.
 8. Tørstofvægtene af toppen antages uafhængige af jordtype og alder.

9. Residualerne antages at have en spredning, der er 95% af konfidensgraden.
10. Effekten af både alder og jordtype antages at være signifikant.
- (c) Undersøg ved hjælp af R-udskriften sidst i opgaven om tørstofvægten af toppen afhænger af jordtype og alder. For det eller de test, du bruger i besvarelsen, skal hypotesen opskrives, og der ønskes præcis angivelse af teststørrelse og P-værdi.
- (d) Formulér en kort og klar konklusion og angiv estimat(er) og konfidensinterval(ler), som du finder passende på baggrund af din konklusion.
- (e) Angiv et estimat og et konfidensinterval som viser forskellen på tørstofvægten mellem alder 2 og alder 3 år. Foretrækker du et sådant estimat og interval angivet for hver jordtype eller fælles for jordtyperne? Begrund svaret.

```

> alder
[1] 1 1 1 1 2 2 2 2 3 3 3 3 1 1 1 1 2 2 2 2 3 3 3 3 1 1 1 1 2 2 2 2 3 3 3 3
> jordtype
[1] A A A A A A A A A A B B B B B B B B B C C C C C C C C C C C C C C C C
Levels: A B C
> top
[1] 4.80 6.99 4.63 7.39 5.21 9.20 4.51 5.98 8.73 5.68 12.42 15.24
[13] 3.14 3.69 2.24 1.08 2.21 3.54 10.23 3.70 6.14 7.40 12.43 8.87
[25] 25.16 10.07 5.18 6.87 11.82 21.13 19.10 13.61 25.49 22.39 20.30 18.92

> alder <- factor(alder)
> model.1 <- lm(top ~ 1)
> model.j <- lm(top ~ jordtype)
> model.a <- lm(top ~ alder)
> model.aj1 <- lm(top ~ alder + jordtype)
> model.aj2 <- lm(top ~ alder*jordtype)
> model.aj3 <- lm(top ~ alder:jordtype - 1)

> summary(model.j)
Call: lm(formula = top ~ jordtype)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.565     1.419    5.330 6.97e-06
jordtypeB   -2.176     2.007   -1.084   0.286
jordtypeC    9.105     2.007    4.536 7.19e-05
---
Residual standard error: 4.917 on 33 degrees of freedom
Multiple R-squared: 0.5187,      Adjusted R-squared: 0.4895
F-statistic: 17.78 on 2 and 33 DF,  p-value: 5.762e-06

```

```

> summary(model.a)
Call: lm(formula = top ~ alder)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.770     1.855   3.649 0.000901
alder2       2.417     2.624   0.921 0.363749
alder3       6.898     2.624   2.629 0.012915
---
Residual standard error: 6.428 on 33 degrees of freedom
Multiple R-squared: 0.1774,      Adjusted R-squared: 0.1275
F-statistic: 3.558 on 2 and 33 DF,  p-value: 0.03988

> summary(model.aj1)
Call: lm(formula = top ~ alder + jordtype)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.460     1.502   2.969 0.005722
alder2       2.417     1.646   1.468 0.152060
alder3       6.898     1.646   4.191 0.000214
jordtypeB   -2.176     1.646  -1.322 0.195807
jordtypeC    9.105     1.646   5.532 4.67e-06
---
Residual standard error: 4.031 on 31 degrees of freedom
Multiple R-squared: 0.696,      Adjusted R-squared: 0.6568
F-statistic: 17.75 on 4 and 31 DF,  p-value: 1.136e-07

> summary(model.aj3)
Call: lm(formula = top ~ alder:jordtype - 1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
alder1:jordtypeA  5.953     2.084   2.856 0.008155
alder2:jordtypeA  6.225     2.084   2.987 0.005937
alder3:jordtypeA 10.518     2.084   5.046 2.69e-05
alder1:jordtypeB  2.538     2.084   1.217 0.233971
alder2:jordtypeB  4.920     2.084   2.361 0.025728
alder3:jordtypeB  8.710     2.084   4.179 0.000275
alder1:jordtypeC 11.820     2.084   5.671 5.06e-06
alder2:jordtypeC 16.415     2.084   7.876 1.82e-08
alder3:jordtypeC 21.775     2.084  10.447 5.51e-11
---
Residual standard error: 4.169 on 27 degrees of freedom

```

Multiple R-squared: 0.9092, Adjusted R-squared: 0.8789
F-statistic: 30.04 on 9 and 27 DF, p-value: 8.344e-12

> anova(model.aj2, model.aj3)

Analysis of Variance Table

Model 1: top ~ alder * jordtype

Model 2: top ~ alder:jordtype - 1

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	27	469.18				
2	27	469.18	0	1.1369e-13		

> anova(model.aj1, model.aj2)

Analysis of Variance Table

Model 1: top ~ alder + jordtype

Model 2: top ~ alder * jordtype

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	31	503.77				
2	27	469.18	4	34.592	0.4977	0.7376

> anova(model.a, model.aj1)

Analysis of Variance Table

Model 1: top ~ alder

Model 2: top ~ alder + jordtype

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	33	1363.34				
2	31	503.77	2	859.57	26.447	1.987e-07

> anova(model.a, model.j)

Analysis of Variance Table

Model 1: top ~ alder

Model 2: top ~ jordtype

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	33	1363.34				
2	33	797.75	0	565.6		

> anova(model.j, model.aj1)

Analysis of Variance Table

Model 1: top ~ jordtype

Model 2: top ~ alder + jordtype

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	33	797.75				
2	31	503.77	2	293.98	9.045	0.0008049

Opgave 2.

Data i denne opgave indeholder 12 målinger af den relative bakterieaktivitet i jordprøver ved forskellige koncentrationer af miljøgiften fenol. Der var tre målinger for hver af koncentrationerne 5 mM, 10 mM, 15 mM og 20 mM.

I begyndelsen af R-programmet nedenfor ses de 12 værdier af den relative bakterieaktivitet, dvs. bakterieaktiviteten ved den givne koncentration i forhold til bakterieaktiviteten uden tilslætning af fenol. Endvidere ses de tilhørende 12 koncentrationer af fenol.

I nedenstående R-udskrift udskrives først datasættet **bakterier**:

```
> bakterier
  rel.aktivitet fenol
 1   0.8840639    5
 2   0.9098648    5
 3   0.8850276    5
 4   0.5414683   10
 5   0.5265748   10
 6   0.5251730   10
 7   0.2732967   15
 8   0.2423707   15
 9   0.2639226   15
10   0.1242736   20
11   0.1145928   20
12   0.1240984   20
```

```
> summary(lm(rel.aktivitet~fenol,data=bakterier))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.098027	0.044003	24.95	2.44e-10
fenol	-0.051744	0.003214	-16.10	1.77e-08

Residual standard error: 0.06223 on 10 degrees of freedom

Multiple R-squared: 0.9629, Adjusted R-squared: 0.9591

F-statistic: 259.3 on 1 and 10 DF, p-value: 1.766e-08

- (a) Opskriv den statistiske model svarende til R-udskriften ovenfor og opskriv kort, gerne i punktform, antagelserne bag modellen.

- (b) Angiv estimerater for alle modellens parametre samt for de koncentrationer af fenol, der (ifølge modellen) ville give en relativ bakterieaktiviteter på henholdsvis 0 og 0.5.
- (c) Den relative bakterieaktivitet ved fenolkoncentrationen nul bør være 1. Stemmer modellen og data med denne oplysning? (Her ønskes en statistisk vurdering, som tager den tilfældige variation i betragtning).

Opgave 3.

Ved undersøgelse af køer fra 6 besætninger fandt man nedenstående antal køer henholdsvis uden og med tegn på infektion, sår eller benlidelser.

Besætning	1	2	3	4	5	6
Antal raske køer	32	32	22	38	36	36
Antal syge køer	12	11	21	13	1	8
Antal køer i alt	44	43	43	51	37	44

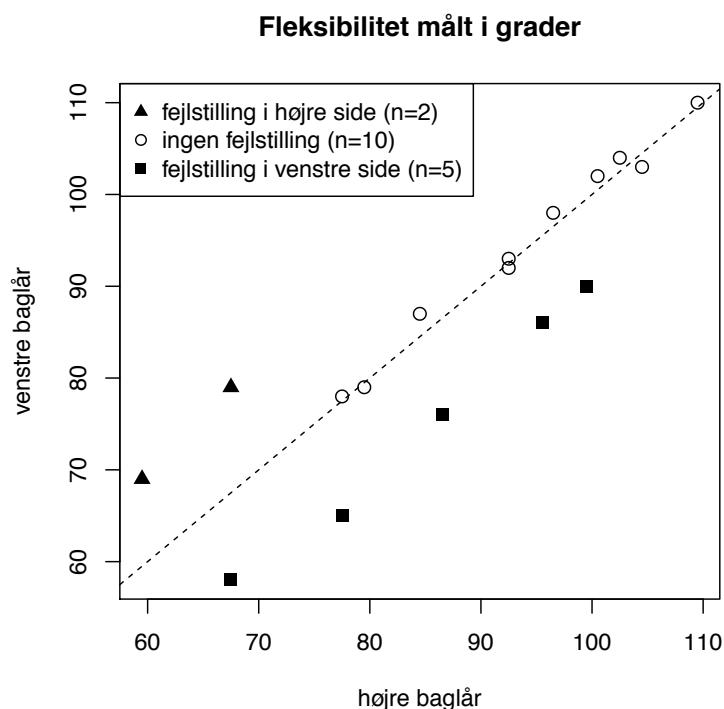
- (a) Hvordan ville du undersøge, om sygeligheden varierer mere mellem besætningerne end hvad, der med rimelighed kan skyldes tilfældigheder? Svaret ønskes angivet i form af formlen for en teststørrelse (test statistic), teststørrelsens fordeling samt en eller nogle få liniers R-kode til udførelse af testet. Du skal således ikke udføre selve testet.
- (b) Antag nu, uanset svaret i første spørgsmål, at sandsynligheden for, at en tilfældigt udvalgt ko er syg, er den samme uanset hvilken besætning, den kommer fra. Angiv et estimat og et konfidensinterval for denne sandsynlighed.
- (c) Antag i dette spørgsmål, at sandsynligheden for, at en tilfældigt udvalgt ko er syg, er 0.15. Hvad er sandsynligheden for, at man ved undersøgelse af 20 (nye) køer ikke finder nogen syge køer?
- (d) Blandt de i alt 262 køer fra de 6 undersøgte besætninger udvælges tilfældigt en af de syge køer. Hvad er sandsynligheden for, at den kommer fra besætning 1? Udover talværdien for denne sandsynlighed skal svaret angive det udfaldsrum og den sandsynlighedsfordeling, som ligger til grund for svaret.

Eksamens i Statistisk Dataanalyse 1 (LMAB10069)
November 2014

Alle hjælpemidler er tilladt, herunder lommeregner og PC. Opgaverne er dog formuleret således at ikke er brug for PC. Der er 5 sider med 3 opgaver og i alt 12 spørgsmål, der alle ønskes besvaret. Den vægt hvormed opgaverne indgår i bedømmelsen er angivet i parentes efter opgave nummeret. Benyttede statistiske modeller skal opskrives, og hypoteser, der testes, skal specificeres både udtrykt ved modellens parametre og i ord. Der er i opgaverne givet beregningshjælp i form af redigerede R udskrifter. Dele heraf skal benyttes, men der kan også være overflødige dele.

Opgave 1. (40%)

Skader af spillere har stor sportslig og økonomisk betydning i moderne top fodbold. For at identificere en mulig årsag til hyppige skader hos nogle fodboldspillere målte en fysioterapeut fleksibiliteten af baglåret på begge ben hos 7 spillere med en såkaldt fejlstilling i bækkenet og hos 10 spillere uden fejlstilling i bækkenet. Fejlstillingen kan være i enten venstre eller højre side af bækkenet, og baglåret i den side hvor fejlstillingen er formodes at være mindre fleksibel end baglåret i den anden side. Figuren nedenfor viser målingerne af fleksibiliteten i venstre og højre baglår hos de 7 spillere med fejlstilling (heraf havde 5 spillere fejlstilling i venstre side) og hos de 10 spillere uden fejlstilling.



Responsvariablen i denne opgave er forskellen mellem fleksibiliteten af venstre og højre baglår. Det kan antages at denne er normalfordelt i alle 3 fejlstillingsgrupper (højre side, ingen, venstre side), og at variansen er den samme i de tre grupper. Følgende R output er lavet ud fra variablen **forsk**, som angiver forskellen mellem fleksibiliteten af venstre og højre baglår, og den kategorisk variabel **fejlstilling** (med niveauerne **højreSide**, **ingen**, **venstreSide**), som angiver hvilken af de 3 fejlstillingsgrupper observationerne kommer fra.

```
> højre <- forskel[fejlstilling=="højreSide"]
> venstre <- forskel[fejlstilling=="venstreSide"]
> ingen <- forskel[fejlstilling=="ingen"]
> mean(højre)
[1] 10.5
> mean(venstre)
[1] -10.3
> mean(ingen)
[1] 0.6
> sd(højre)
[1] 1.414214
> sd(venstre)
[1] 1.30384
> sd(ingen)
[1] 1.197219
```

Besvar følgende spørgsmål:

- Er forskellen i fleksibiliteten mellem venstre og højre baglår forskellig i populationen af spillere med fejlstilling i venstre side af bækkenet og i populationen af spillere uden fejlstilling i bækkenet?
- Lav et 95% konfidensinterval for forskellen mellem fleksibiliteten af venstre og højre baglår i populationen af spillere uden fejlstilling i bækkenet. Kan man antage at venstre og højre baglår er lige fleksible i denne population?
- Vil det være usædvanligt, hvis en spiller uden fejlstilling i bækkenet havde en fleksibilitet på 91 grader i venstre baglår og 94 grader i højre baglår?
- Betrægt nu de 7 spillere med fejlstilling i bækkenet. Formulér den statistiske hypotese at effekten af fejlstilling i bækkenet på fleksibiliteten af baglåret i den berørte side ikke afhænger af om fejlstillingen er i venstre eller højre side i termer af middelværdi parametrene for de tilsvarende populationer. Og lav et statistisk test af denne hypotese.
- Sammenlign forskellen i fleksibiliteten mellem baglåret i den berørte side og den anden side i populationen af spillere med fejlstilling (enten i venstre eller højre side af bækkenet) med forskellen i fleksibiliteten mellem venstre og højre baglår i populationen af spillere uden fejlstilling. Til det formål kan du bruge R output'et, som findes på næste side.

```
> reduktion <- c(venstre,-højre)
> t.test(reduktion,ingen,var.equal=TRUE)
```

Two Sample t-test

```
data: reduktion and ingen
t = -18.4615, df = 15, p-value = 1.003e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-12.222187 -9.692098
sample estimates:
mean of x mean of y
-10.35714   0.60000
```

Opgave 2. (20%)

I denne opgave analyseres et andet aspekt af undersøgelsen beskrevet i opgave 1. Ialt undersøges 85 fodboldspillere for fejlstilling i bækkenet. Heref havde 5 og 2 spillere fejlstilling i henholdsvis venstre og højre side af bækkenet, og 78 spillere var uden fejlstilling i bækkenet. Antag at de undersøgte fodboldspillere er repræsentative for populationen af mænd mht. fejlstilling i bækkenet.

Besvar følgende spørgsmål:

- Giver stikprøven anledning til at konkludere, at fejlstillingen ikke forekommer lige hyppigt i venstre og højre side af bækkenet i populationen af mænd med fejlstilling?
- Bestem et 95% konfidensinterval for hyppigheden af fejlstilling i bækkenet i populationen af mænd.

Opgave 3. (40%)

En forsker undersøgte hvorledes unge baobab træer vokser under varierende grader af vandmangel. I denne opgave analyseres hvorledes træernes volumen (beregnet ud fra målinger af diameter og højde) ændrer sig i vækstperioden fra januar til august måned når træerne har haft høj, middel, eller lavt vandningsniveau (100%, 75%, eller 50% vand).

Data består af målinger på 460 træer, som vi antager findes i R i data frame'en **baobab** med variablerne **volumen**, **vand** og **varighed**. Her er **volumen** responsevariablen, **vand** er en kategorisk forklarende variabel med niveauerne 100%, 75% og 50%, som angiver vandningsregimet, og **varighed** er en kontinuert forklarende variabel som angiver hvor lang tid træerne har vokset under de forskellige vandningsregimer. De tre vandningsregimer blev igangsat i januar måned, hvormed **varighed=1** svarer til februar måned og **varighed=7** svarer til august måned.

Besvar følgende spørgsmål:

- (a) Det viser sig, at logaritmen af volumen kan modelleres via en lineær regression mht. varighed, men hvor både intercept og hældning afhænger af vandingsniveauet. Angiv R kode som kan bruges til at estimere og validere denne model.
- (b) Antag at modellen lavet i (a) findes i lm-objektet `m1` i R. Inden vandingsbehandlingerne starter bør der ikke være forskel på træernes volumen i de 3 vandingsgrupper. Formulér dette udsagn som en nul hypotese om parametrene i `m1`, og brug R output'et nedenfor til at diskutere om data er i overensstemmelse med denne nul hypotese.

```
> summary(m1)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.88702	0.11594	42.151	< 2e-16 ***
vand50%	-0.08242	0.16359	-0.504	0.615
vand75%	-0.06377	0.16494	-0.387	0.699
varighed	0.32731	0.01665	19.654	< 2e-16 ***
vand50%:varighed	-0.17488	0.02361	-7.407	6.35e-13 ***
vand75%:varighed	-0.10656	0.02358	-4.519	7.92e-06 ***

Residual standard error: 0.699 on 454 degrees of freedom

Multiple R-squared: 0.6516, Adjusted R-squared: 0.6477

F-statistic: 169.8 on 5 and 454 DF, p-value: < 2.2e-16

Vi estimerer nu en model, hvor der er et fælles intercept for de 3 vandingsgrupper, men hvor hældningen mht. varighed afhænger af vandingsgruppen. R output fra denne model er givet nedenfor. Besvar følgende spørgsmål:

- (c) Angiv et 95% konfidensinterval for forholdet mellem volumen af baobab træer i august måned (dvs. 7 måneder efter forsøgets start) når de henholdsvis har fået 50% og 100% vand siden januar måned.

```
> m2 <- lm(log(volumen)~varighed+vand:varighed,data=baobab)
```

```
> summary(m2)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.83830	0.06697	72.25	<2e-16 ***
varighed	0.33339	0.01172	28.45	<2e-16 ***
varighed:vand50%	-0.18522	0.01163	-15.93	<2e-16 ***
varighed:vand75%	-0.11453	0.01136	-10.09	<2e-16 ***

Residual standard error: 0.6976 on 456 degrees of freedom

Multiple R-squared: 0.6513, Adjusted R-squared: 0.6491

F-statistic: 284 on 3 and 456 DF, p-value: < 2.2e-16

Antag, at væksten af baobab træer fra januar til august måned kan beskrives ved ligningerne:

$$\text{Hvis vand}=50\%: \quad \log(\text{volumen}_i) = 4.8 + 0.15 \cdot \text{varighed}_i + e_i,$$

$$\text{Hvis vand}=75\%: \quad \log(\text{volumen}_i) = 4.8 + 0.22 \cdot \text{varighed}_i + e_i,$$

$$\text{Hvis vand}=100\%: \quad \log(\text{volumen}_i) = 4.8 + 0.33 \cdot \text{varighed}_i + e_i,$$

hvor $\log(x)$ angiver den naturlige logaritme, hvor volumen er målt i cm^3 , hvor varighed er antal måneder siden januar, og hvor fejleddene e_1, \dots, e_{460} er uafhængige og normalfordelte med middelværdi 0 og spredning 0.7.

Besvar følgende spørgsmål:

- (d) Hvad er sandsynligheden for at et baobab træ der har fået 75% vand siden januar måned har en volumen mindre end 150 cm^3 i maj måned?
- (e) Betragt volumen af to baobab træer i august måned, som har fået henholdsvis 50% og 100% vand siden januar måned. Hvad er sandsynligheden for at det træ som har fået 50% vand har den største volumen?

Reeksamen i Statistisk Dataanalyse 1 (LMAB10069)

Januar 2015

Alle hjælpemidler er tilladt, herunder lommeregner og PC. Opgaverne er dog formuleret således, at der ikke er brug for PC til beregningerne. Der er 5 sider med 3 opgaver og i alt 12 delspørgsmål, der alle ønskes besvaret. Delspørgsmålene indgår med samme vægt i bedømmelsen af besvarelsen. Benyttede statistiske modeller skal opskrives, og hypoteser, der testes, skal specificeres både udtrykt ved modellens parametre og i ord. Der er i opgaverne givet beregningshjælp i form af redigerede R udskrifter. Dele heraf skal benyttes, men der kan også være overflødige dele.

Opgave 1. (4 delspørgsmål)

Satisfaction With Life Index (SWL) er et forsøg på at kvantificere befolkningens tilfredshed med livet, eller lykke om man vil, i verdens lande (jo større SWL jo højere "lykke"). Danskerne er adskillige gange blevet kåret som verdens lykkeligste folk på SWL skalaen. I denne opgave undersøges om der er en sammenhæng mellem SWL og skattetryk. Data består af sammenhørende målinger af land, SWL, selskabsskat, indkomstskat, og omsætningsafgift for 55 lande. De tre skattevariable betegnes ved hhv. **selskab**, **indkomst** og **VAT**, de måles i procent, og de angiver den øvre grænse for pågældende skat i de enkelte lande. Antag at data findes i en dataframe ved navn **happy.tax**, og betragt følgende R output:

```
> summary(happy.tax)
      land        SWL        selvkab      indkomst       VAT
Argentina: 1   Min.   :120.0   Min.   :10.00   Min.   :10.00   Min.   : 0.00
Australia: 1   1st Qu.:188.3   1st Qu.:20.00   1st Qu.:26.25   1st Qu.:12.50
Austria   : 1   Median :220.0   Median :25.00   Median :35.00   Median :18.00
Belgium   : 1   Mean    :212.2   Mean    :25.25   Mean    :34.28   Mean    :15.87
Brazil    : 1   3rd Qu.:243.3   3rd Qu.:30.00   3rd Qu.:41.75   3rd Qu.:20.00
Bulgaria  : 1   Max.    :273.4   Max.    :40.00   Max.    :59.00   Max.    :25.00
(Other)   :49
> m1 <- lm(SWL~indkomst,data=happy.tax)
> summary(m1)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	145.7468	11.9497	12.197	< 2e-16 ***
indkomst	1.9397	0.3301	5.876	2.85e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.45 on 53 degrees of freedom
Multiple R-squared: 0.3945, Adjusted R-squared: 0.383
F-statistic: 34.52 on 1 and 53 DF, p-value: 2.852e-07

Besvar følgende spørgsmål:

- (a) Undersøg om der er en sammenhæng mellem SWL og indkomstskatteprocenten i de 55 undersøgte lande, og beskriv i givet fald denne sammenhæng.
- (b) Angiv et 95% konfidensinterval for SWL i lande med indkomstskat på 59 procent. Du må i den forbindelse gerne bruge følgende R beregninger:

```
> mean(happy.tax$indkomst)
[1] 34.28182

> var(happy.tax$indkomst)*(55-1)    # Dette er SS_x værdien
[1] 7428.382

> qt(0.975,df=55-2)
[1] 2.005746
```

- (c) I hvilket interval vil man med 95% sandsynlighed forvente at finde SWL for et land med indkomstskat på 35 procent?
- (d) Brug modellen m2 fra R output'et nedenfor til at beskrive om, og i givet fald hvordan, variablerne **selskab**, **indkomst** og **VAT** påvirker SWL i de 55 undersøgte lande.

```
> m2 <- lm(SWL ~ selskab+indkomst+VAT,data=happy.tax)
> summary(m2)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)							
(Intercept)	173.2095	17.0829	10.139	8.03e-14 ***							
selskab	-0.6602	0.5881	-1.123	0.2669							
indkomst	2.2819	0.3646	6.258	8.06e-08 ***							
VAT	-1.4195	0.6023	-2.357	0.0223 *							

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'.'	0.1	'	1

Residual standard error: 27.38 on 51 degrees of freedom
Multiple R-squared: 0.4603, Adjusted R-squared: 0.4286
F-statistic: 14.5 on 3 and 51 DF, p-value: 5.925e-07

Opgave 2. (5 delspørgsmål)

Mange kvinder oplever forhøjet blodtryk under graviditeten. I et forsøg blev det undersøgt om aspirin kan afhjælpe dette.¹ 65 gravide kvinder med risiko for forhøjet blodtryk under graviditeten blev inddelt i to grupper. Den første gruppe fik dagligt 100 mg aspirin i den sidste tredjedel af graviditeten, mens den anden gruppe fik placebobehandling i form af kalktabletter. Resultatet fra forsøget er givet i tabellen nedenfor.

Forhøjet blodtryk	Gruppe		Total
	Aspirin	Placebo	
Ja	4	11	15
Nej	30	20	50
Total	34	31	65

Besvar følgende spørgsmål:

- (a) Kan aspirin bruges til at forebygge forhøjet blodtryk under graviditeten?

I en anden dataindsamling fra 351 kvinder undersøgtes sammenhængen mellem den gravides skostørrelse og risikoen for at der blev foretaget kejsersnit under fødslen. Resultatet fra dette forsøg er givet i tabellen nedenfor, hvor “≤ 36.5” betyder skostørrelse 36.5 eller mindre, og “≥ 40” betyder skostørrelse 40 eller større.

Kejsersnit	Skostørrelse						Total
	≤ 36.5	37	37.5	38	39	≥ 40	
Ja	5	7	6	7	8	10	43
Nej	17	28	36	41	46	140	308
Total	22	35	42	48	54	150	351

Besvar følgende spørgsmål:

- (b) Hvorfor kan man ikke benytte et chi-kvadrat test på ovenforstående 2×6 tabel til at afgøre, om der er en sammenhæng mellem skostørrelse og risiko for kejsersnit?
- (c) Slå skostørrelsесgrupperne sammen to og to, således at der dannes en 2×3 tabel med kolonnerne “≤ 37”, “37.5 eller 38”, og “≥ 39”. Undersøg om der er en sammenhæng mellem skostørrelse og risiko for at få kejsersnit, og beskriv i givet fald denne sammenhæng.
- (d) Angiv et 95% konfidensinterval for risikoen for at få kejsersnit for kvinder, der har skostørrelse 37 eller mindre.
- (e) På en fødegang var der 10 fødende kvinder, hvoraf 2 havde skostørrelse 37 eller mindre, 3 havde skostørrelse 37.5 eller 38, og 5 havde skostørrelse 39 eller større. Angiv et estimat for sandsynligheden for at en tilfældig udvalgt kvinde blandt disse 10 kvinder fik kejsersnit.

¹Aspirin har andre bivirkninger som gør, at aspirin generelt set ikke bør bruges under graviditeten.

Opgave 3. (3 delspørgsmål)

En producent af dyrefoder undersøgte, hvordan vitamin koncentrationen i det producerede foder afhænger af mængden af et tilsætningsstof, og temperaturen hvorved tilsætningsstoffet blev blandet i foderet. Der blev afprøvet 4 forskellige mængder af tilsætningsstoffet (stigende fra niveau *A* til niveau *D*) ved 3 forskellige temperaturer (stigende fra niveau *I* til niveau *III*), og for hver kombination blev der lavet 2 målinger. Der er altså i alt 24 målinger af vitamin koncentration, som er angivet i tabellen nedenfor.

temperatur	mængde tilsætningsstof			
	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>
<i>I</i>	33 37	50 56	72 64	77 73
<i>II</i>	60 64	79 73	80 88	79 87
<i>III</i>	46 50	72 58	72 80	81 77

Betrægt følgende R output, hvor dataframe'en **foder** indeholder de 24 sammenhørende målinger af vitamin koncentration, mængde tilsætningsstof og temperatur.

```
> m1 <- lm(vitamin~temperatur*tilsætning,data=foder)
> drop1(m1,test="F")
Single term deletions
```

Model:

```
vitamin ~ temperatur * tilsætning
          Df Sum of Sq    RSS      AIC F value Pr(>F)
<none>                 302.00 84.777
temperatur:tilsætning  6     209.67 511.67 85.431  1.3885 0.2953
```

```
> m2 <- lm(vitamin~temperatur+tilsætning,data=foder)
> summary(m2)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	39.083	2.666	14.661	1.89e-11 ***
temperaturII	18.500	2.666	6.940	1.74e-06 ***
temperaturIII	9.250	2.666	3.470	0.00273 **
tilsætningB	16.333	3.078	5.306	4.81e-05 ***
tilsætningC	27.667	3.078	8.988	4.49e-08 ***
tilsætningD	30.667	3.078	9.963	9.46e-09 ***

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'
	0.1 ' '	1		

```
Residual standard error: 5.332 on 18 degrees of freedom
Multiple R-squared:  0.9044, Adjusted R-squared:  0.8779
F-statistic: 34.07 on 5 and 18 DF,  p-value: 1.433e-08
```

Besvar følgende spørgsmål:

- (a) Opskriv den statistiske model hørende til $m1$, forklar hvilken hypotese der er blevet testet i denne model, og angiv konklusionen på det udførte test.
- (b) Brug den såkaldte *additive model*, som findes i $m2$, til at beregne et 95% konfidensinterval for vitamin koncentrationen når mængden af tilsætningsstof er på niveau A , og blandingen sker ved temperatur på niveau I .
- (c) Brug estimerater og standard errors fra den *additive model* til at lave et t-test for hypotesen, at vitamin koncentrationen er den samme når mængden af tilsætningsstof er på niveau C og D . På forhånd forventede producenten, at der findes et niveau for mængden af tilsætningsstoffet, således at der ikke er yderligere effekt af at tilsætte mere af tilsætningsstoffet ud over dette niveau. Understøtter data denne forhåndsforventning?

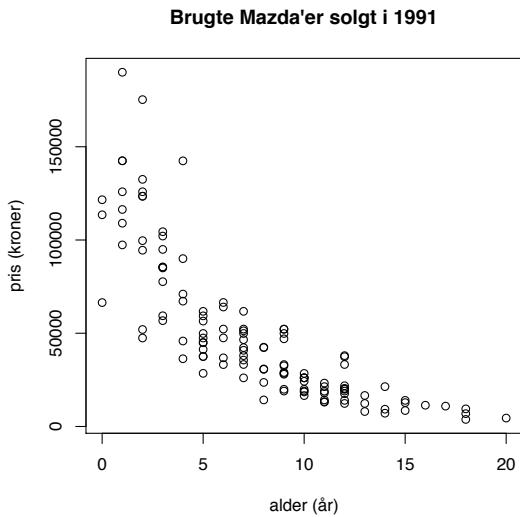
Eksamensopgave i Statistisk Dataanalyse 1 (LMA10069U)

November 2015

Alle hjælpemidler er tilladt, herunder lommeregner og PC. Opgaverne er dog formuleret således at ikke er brug for PC. Der er 7 sider med 3 opgaver og i alt 14 delspørgsmål, der alle ønskes besvaret. I bedømmelsen indgår alle delspørgsmål med samme vægt. Benyttede statistiske modeller skal opskrives, og hypoteser, der testes, skal specificeres både udtrykt ved modellens parametre og i ord. Der er i opgaverne givet beregningshjælp i form af redigerede R udskrifter. Dele heraf skal benyttes, men der kan også være overflodige dele.

Opgave 1.

Nedenstående figur viser sammenhørende værdier af *pris* (i danske kroner) og *alder* (i år) for 121 brugte biler af mærket Mazda solgt i 1991.

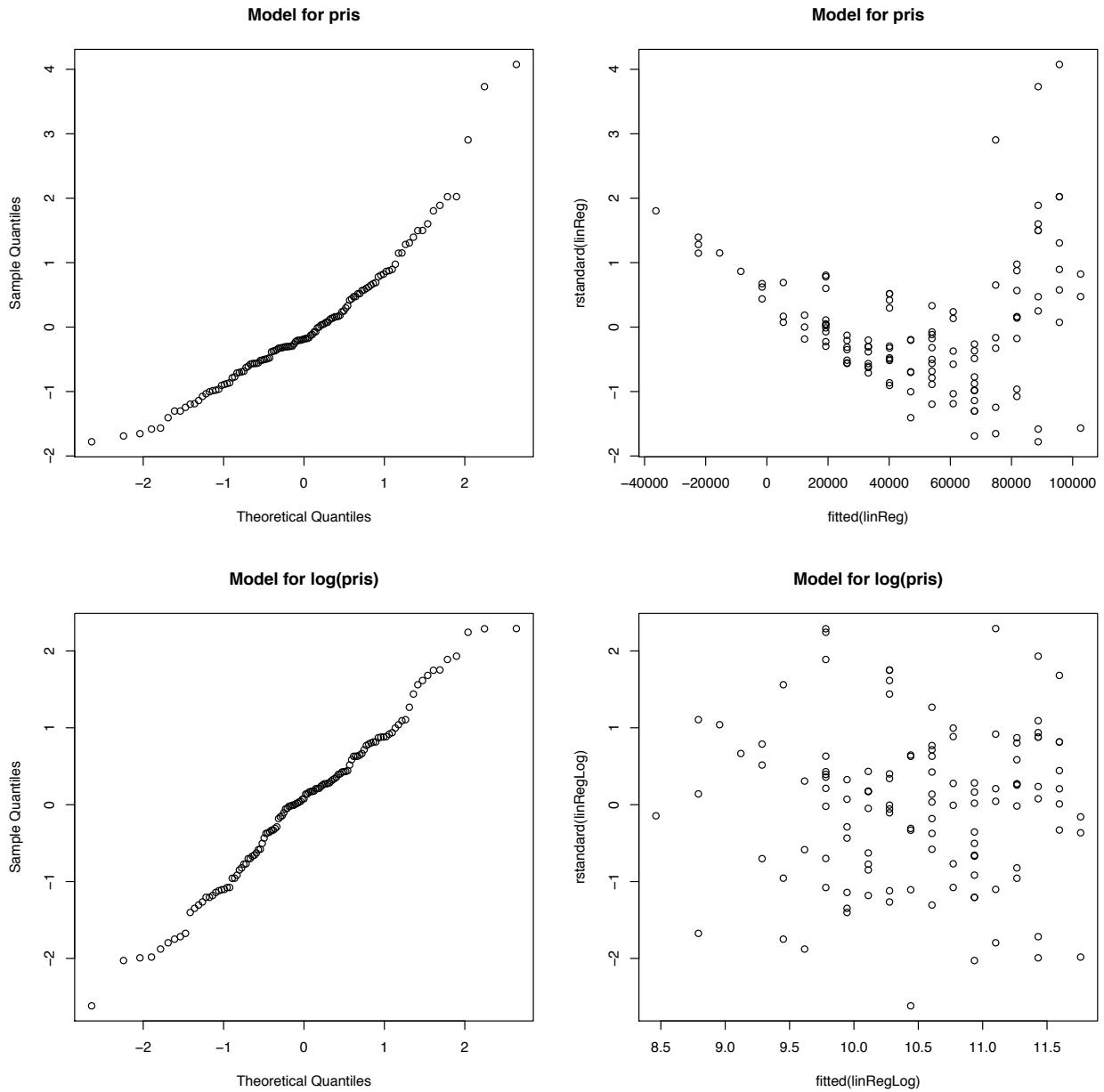


Antag, at disse data er tilgængelige i R i en data frame ved navn `mazda` med variablerne `pris` og `alder`. Vi ønsker at beskrive prisen som funktion af bilens alder ved hjælp af en af følgende to lineære regressioner:

```
> linReg <- lm(pris~alder,data=mazda)
> linRegLog <- lm(log(pris)~alder,data=mazda)
```

For at se om en af disse modeller er statistisk valid laves model validering. Det grafiske output fra nedenstående R kode findes på næste side af opgavesættet.

```
> qqnorm(rstandard(linReg),main="Model for pris")
> plot(fitted(linReg),rstandard(linReg),main="Model for pris")
> qqnorm(rstandard(linRegLog),main="Model for log(pris)")
> plot(fitted(linRegLog),rstandard(linRegLog),main="Model for log(pris)")
```



Besvar følgende spørgsmål, eventuelt ved brug af R output som findes i slutningen af denne opgave:

- (a) Redegør for hvilken af modellerne `linReg` og `linRegLog` der bedst opfylder model antagelserne (svaret skal begrundes). Du skal bruge denne model ved besvarelse af de følgende delspørgsmål.

- (b) Opskriv den valgte model. Og angiv estimerater for alle model parametrene, og 95% konfidensintervaller for de parametre der beskriver middelværdien.
- (c) Lav et 95% konfidensinterval for prisen på en brugt Mazda der er 5 år gammel.
- (d) Ville det være usædvanligt hvis en brugt Mazda der er 5 år gammel kostede 80000 kroner? Svaret skal begrundes.
- (e) Angiv estimatet for alderen på en bil, hvor den forventede pris er halvt så stor som den forventede pris på en 0 år gammel brugt bil.

R output (lidt redigeret) til opgave 1:

```
> mean(mazda$alder)
[1] 7.628099

> sd(mazda$alder)
[1] 4.566421

> summary(linReg)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 102606.3     4161.9   24.65 <2e-16 ***
alder       -6944.5      468.6  -14.82 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23440 on 119 degrees of freedom
Multiple R-squared:  0.6485, Adjusted R-squared:  0.6456
F-statistic: 219.6 on 1 and 119 DF,  p-value: < 2.2e-16

> summary(linRegLog)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.760901  0.059744 196.86 <2e-16 ***
alder       -0.164965  0.006727 -24.52 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3365 on 119 degrees of freedom
Multiple R-squared:  0.8348, Adjusted R-squared:  0.8334
F-statistic: 601.3 on 1 and 119 DF,  p-value: < 2.2e-16
```

Opgave 2.

Hvis hastigheden hvormed en lyskilde blinker (målt i antal blink per sekund) bliver tilstrækkeligt høj, så vil et menneske ikke længere kunne se at lyset blinker, men vil i stedet for opfatte lyset som en konstant lyskilde. Den *kritiske frekvens* for en given person defineres som det højeste antal blink per sekund hvorved personen stadigvæk kan opfatte, at en blinkende lyskilde blinker.

En eksperimentel psykolog undersøgte om den *kritiske frekvens* afhænger af personens øjenfarve. Nedenstående tabel indeholder den *kritiske frekvens* for 19 personer inddelt efter forsøgspersonernes øjenfarve:

Øjenfarve			
Blå	Brun	Grøn	
25.7 27.2 29.9 28.5	26.8 27.9 23.7 25.0	26.4 24.2 28.0 26.9	
29.4 28.3	26.3 24.8 25.7 24.5	29.1	

Nedenfor findes output fra en R analyse af dette datasæt. Man kan uden begrundelse antage, at den benyttede model er statistisk valid.

```
> model <- lm(frekvens~øjenfarve)
```

```
> drop1(model,test="F")
```

```
Single term deletions
```

Model:

```
frekvens ~ øjenfarve
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>		38.310	19.324			
øjenfarve	2	22.997	61.307	24.258	4.8023	0.02325 *

```
---
```

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ',' 1
```

```
> summary(model)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	28.1667	0.6317	44.588	< 2e-16 ***
øjenfarveBrun	-2.5792	0.8357	-3.086	0.00708 **
øjenfarveGrøn	-1.2467	0.9370	-1.331	0.20200

```
---
```

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ',' 1
```

Residual standard error: 1.547 on 16 degrees of freedom

Multiple R-squared: 0.3751, Adjusted R-squared: 0.297

F-statistic: 4.802 on 2 and 16 DF, p-value: 0.02325

Besvar følgende spørgsmål:

- Opskriv den statistisk model der er blevet anvendt. Kan man ud fra datasættet konkludere, at øjenfarven har betydning for en persons *kritiske frekvens*? Svaret skal begrundes.
- I hvilket interval ligger den *kritiske frekvens* for 95% af mennesker med blå øjne?
- Angiv et 95% konfidensinterval for forskellen mellem gennemsnittene for den *kritiske frekvens* i populationerne af personer med henholdsvis brune og grønne øjne.
- Mellem hvilke øjenfarver er der signifikant forskel i populationsgennemsnittene af den *kritiske frekvens*, når der tages højde for multipel testing via en Bonferroni korrektion? Svaret skal begrundes.

Opgave 3.

En underviser på et indledende statistik kursus lavede et simpelt eksperiment med sine studerende. De studerende skulle først tage deres egen puls. Derefter blev de studerende bedt om at kaste *plat og krone* med en mønt. De studerende som fik *krone* skulle løbe på stedet i et minut. Og de studerende som fik *plat* skulle blot blive siddende på deres pladser. Til sidst skulle de studerende tage deres egen puls igen. For hver studerende blev gruppen (*løb, hvile*) og ændringen i pulsen registreret.

Dette forsøg blev gennemført på statistik kurset i årene 1993 og 1995, dog med den forskel at underviseren slog *plat og krone* for de studerende i 1995. Grunden til dette var, at underviseren havde mistanke om, at nogle af de studerende fra 1993 årgangen havde valgt at blive siddende på deres pladser, selv om de havde slæbt *krone*. Antallet af studerende i de to grupper på de to årgange er angivet i 2-vejs antalstabellen nedenfor:

Årgang	Gruppe	
	Løb	Hvile
1993	7	18
1995	9	12

Lad $p_{løb}$ være sandsynligheden for at en studerende fra 1993 årgangen bliver registreret i gruppen, der har løbet på stedet i et minut. Vi opstiller nu følgende nul hypotese

$$H_0: p_{løb} = 0.5$$

Besvar følgende spørgsmål:

- Forklar hvorfor H_0 er en relevant nul hypotese.

- (b) Der er forskellige måder, hvorpå man kan teste H_0 mod alternativet at $p_{løb} \neq 0.5$. Skriv (uden begrundelse) i din besvarelse en indrammet linje med numrene på de af nedenstående tests det er meningsfyldt at bruge.
1. Et chi-kvadrat test for uafhængighed i 2-vejs antalstabellen.
 2. Et chi-kvadrat test for goodness-of-fit for kendte sandsynligheder i en 1-vejs antalstabbel.
 3. Et binomialtest.
- (c) Udfør et af de hypotese tests du har valgt i spørgsmål (b). Er der noget der tyder på, at nogle af de studerende fra årgang 1993 valgte at blive siddende selv om de havde slået krone?

Vi vælger nu kun at bruge puls målingerne fra årgang 1995. I R output, som findes i slutningen af denne opgave, angiver vektorerne x og y pulsændringerne for de 9 og 12 studerende fra årgang 1995 som henholdsvis $løb$ på stedet eller var i hvile. Besvar følgende spørgsmål:

- (d) Hvilken af følgende tre statistiske modeller bør bruges ved beregningen af et 95% konfidensinterval for forskellen mellem pulsændringen for grupperne der henholdsvis $løb$ på stedet og var i hvile. Svaret skal begrundes.
1. Modellen for et t-test for to parrede stikprøver.
 2. Modellen for et t-test for to uparrede stikprøver med ens varians.
 3. Modellen for et t-test for to uparrede stikprøver med forskellig varians.
- (e) Opskriv den statistiske model du har valgt i spørgsmål (d), og angiv det tilhørende 95% konfidensinterval for forskellen mellem pulsændringen for grupperne der henholdsvis $løb$ på stedet og var i hvile.

R output til opgave 3:

```
> x <- c(90,64,33,57,55,94,44,49,60)
> y <- c(-3,-2,0,0,2,6,2,0,-6,-8,0,-2)
> mean(x)
[1] 60.66667
> mean(y)
[1] -0.9166667
> sd(x)
[1] 20.02498
> sd(y)
[1] 3.704011
```

```
> t.test(x,y,var.equal=TRUE)
```

Two Sample t-test

```
data: x and y
t = 10.504, df = 19, p-value = 2.371e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 49.31191 73.85475
sample estimates:
 mean of x  mean of y
60.6666667 -0.9166667
```

```
> t.test(x,y,var.equal=FALSE)
```

Welch Two Sample t-test

```
data: x and y
t = 9.1098, df = 8.4118, p-value = 1.219e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 46.12637 77.04029
sample estimates:
 mean of x  mean of y
60.6666667 -0.9166667
```

Reeksamen i Statistisk Dataanalyse 1 (LMAB10069U)
Februar 2016

Alle hjælpemidler er tilladt, herunder lommeregner og PC. Opgaverne er dog formuleret således at der ikke er brug for PC. Om ønsket må opgaven besvares med blyant. Der er 6 sider med 3 opgaver og i alt 14 delspørgsmål, der alle ønskes besvaret. I bedømmelsen indgår alle delspørgsmål med samme vægt. Der er i opgaverne givet beregningshjælp i form af redigerede R udskrifter. Dele heraf skal benyttes, men der kan også være overflødige dele.

Opgave 1.

Denne opgave omhandler 579 gravide kvinder, hvoraf 59 fødte børn med lav fødselsvægt (som defineres til at være under 2500 gram). Vi vil undersøge om rygning og antallet af besøg hos jordemoderen under graviditeten er risikofaktorer for lav fødselsvægt. Besvar nedenstående spørgsmål:

- (a) Følgende 2×2 -tabel viser krydstabuleringen af fødselsvægt (lav/ikke-lav) og om moderen ryger (nej/ja). Undersøg på baggrund af denne tabel om rygning er en risikofaktor for lav fødselsvægt.

Fødselsvægt	Ryger	
	nej	ja
≤ 2500 gram	29	30
> 2500 gram	345	175

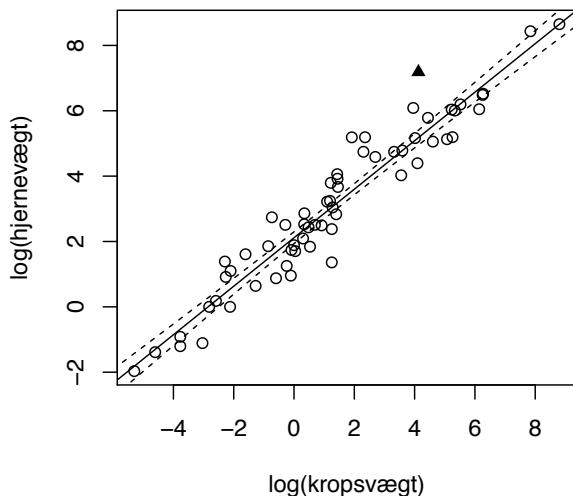
- (b) Angiv et 95% konfidensinterval for forskellen i risikoen for at føde et barn med lav fødselsvægt mellem rygere og ikke-rygere.
- (c) Følgende 2×4 -tabel viser krydstabuleringen af fødselsvægt (lav/ikke-lav) og antal jordemoderbesøg under graviditeten. Her betyder “3+” at moderen havde 3 eller flere besøg hos jordemoderen. Angiv det forventede antal kvinder for hver af de otte kombinationer af fødselsvægt og antal jordemoderbesøg under antagelse af, at der ikke er nogen sammenhæng mellem fødselsvægt og antal jordemoderbesøg.

Fødselsvægt	Jordemoderbesøg			
	0	1	2	3+
≤ 2500 gram	36	11	7	5
> 2500 gram	257	146	91	26

- (d) Angiv den 2×3 -tabel, der viser krydstabuleringen mellem fødselsvægt (lav/ikke-lav) og antallet af jordemoderbesøg, hvor antallet af besøg er grupperet i "0", "1", "2+". Her betyder "2+" at moderen havde 2 eller flere besøg hos jordemoderen. Hvilken krydstabulering ville du bruge til at undersøge om der er en sammenhæng mellem fødselsvægt og antal jordemoderbesøg? Du skal ikke udføre det statistiske test, men dit svar skal begrundes.
- (e) I en mødregruppe er der 7 rygere og 13 ikke-rygere. Bestem med udgangspunkt i tabelerne ovenfor det forventede antal af disse 20 mødres børn, som havde en lav fødselsvægt.

Opgave 2.

Har *mennesket* en usædvanlig stor hjerne? En menneskehjerne vejer 1320 gram, mens en hjerne fra en afrikansk elefant vejer 5712 gram. Så i den forstand er svaret *nej*. Men hvis den naturlige logaritme af hjernevægten (målt i *gram*) tegnes mod den naturlige logaritme af kropsvægten (målt i *kg*) for 62 forskellige landlevende pattedyrsarter^{1,2}, så fås følgende figur:



Punktet tegnet med en udfyldt trekant er *mennesket*, og de 61 punkter tegnet med åbne cirkler er 61 andre pattedyrsarter. Regressionslinjen og de tilhørende 95% konfidensintervaller på figuren er for en lineær regression af $\log(\text{hjernevægt})$ på $\log(\text{kropsvægt})$ for de 61 øvrige

¹Disse vægtangivelser er estimeret for middelværdien i de forskellige artspopulationer. Men i denne opgave vil krops- og hjernevægt blive brugt som observationen af to variable på hver enkelt pattedyrsart.

²Ifølge bogen *Mammal Species of the World* var der 5416 kendte pattedyrsarter i 2006. Så 62 pattedyrsarter udgør kun en lille del af samtlige kendte arter.

arter (altså uden *mennesket*). I besvarelsen af de følgende spørgsmål kan det antages, at denne model er statistisk valid. R output fra modellen findes i slutningen af denne opgave.

- Opskriv den brugte statistiske model. Angiv estimerer for alle model parametrene, og 95% konfidensintervaller for de parametre der beskriver middelværdien.
- Brug den statistiske model til at beregne et 95% konfidensinterval for hjernevægten for en pattedyrsart med kropsvægt på *62 kg*.
- Krops- og hjernevægten af *mennesket* er henholdsvis *62 kg* og *1320 gram*. Kan man sige, at *mennesket* har en usædvanlig stor hjerne? Svaret skal begrundes.
- Blandt de 61 øvrige pattedyrsarter i datasættet er *rhesusaben* den art, der har den største hjernevægt i forhold til sin kropsvægt. Kropsvægten for en *rhesusabe* er *6,8 kg*. Hvor mange gange tungere end en rhesusabehjerne skulle en menneskehjerne være ifølge den statistiske model? Angiv også et 95% konfidensinterval for denne kvotient.

R output (lidt redigeret) til opgave 2: Bemærk, at *mennesket* er observation nr. 34 i data frame'en *pattedyr*, hvormed denne observation er fjernet i beregningerne nedenfor.

```
> summary(pattedyr)
      art          kropsvægt        hjernevægt
Africanelephant    : 1   Min.   : 0.005   Min.   : 0.14
Africangiantpouchedrat: 1   1st Qu.: 0.600   1st Qu.: 4.25
ArcticFox           : 1   Median : 3.342   Median : 17.25
Arcticgroundsquirrel: 1   Mean    : 198.790  Mean    : 283.13
Asianelephant       : 1   3rd Qu.: 48.203   3rd Qu.: 166.00
Baboon              : 1   Max.    :6654.000  Max.    :5712.00
(Other)             :56
```

```
> mean(log(pattedyr$hjernevægt[-34]))
[1] 3.073883
```

```
> sd(log(pattedyr$hjernevægt[-34]))
[1] 2.409975
```

```
> mean(log(pattedyr$kropsvægt[-34]))
[1] 1.291808
```

```
> sd(log(pattedyr$kropsvægt[-34]))
[1] 3.128045
```

```
> linreg <- lm(log(hjernevægt)^log(kropsvægt), data=pattedyr[-34,])
```

```

> summary(linreg)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.11500   0.09030 23.42   <2e-16 ***
log(kropsvægt) 0.74228   0.02687 27.62   <2e-16 ***  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

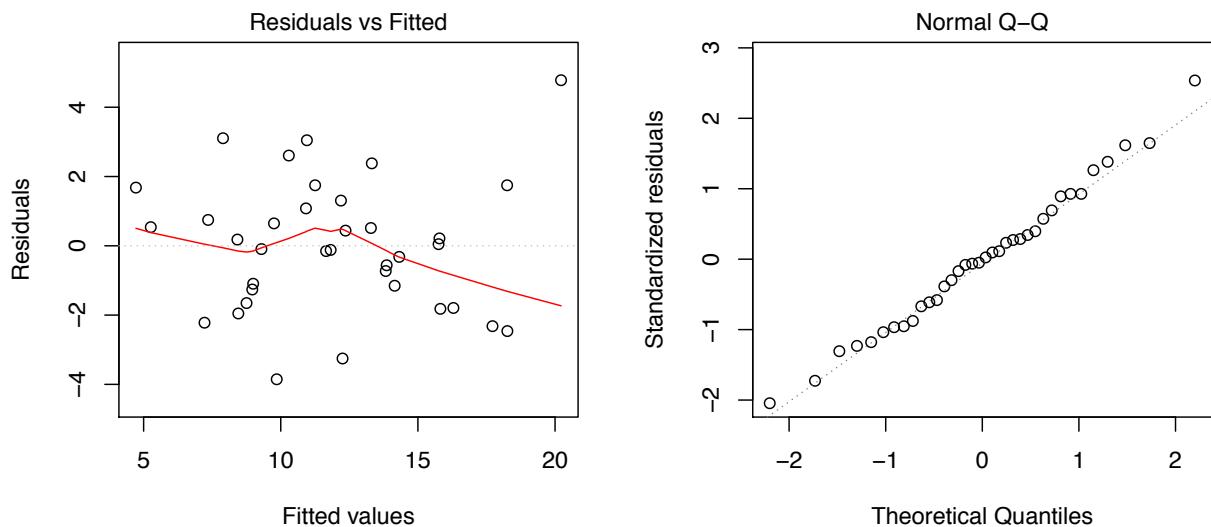
Residual standard error: 0.6511 on 59 degrees of freedom
Multiple R-squared:  0.9282, Adjusted R-squared:  0.927 
F-statistic:    763 on 1 and 59 DF,  p-value: < 2.2e-16

```

Opgave 3.

I folketroen har fuldmånen en ond og mørk kraft der kan styre vores adfærd. Men er der noget om snakken? Datasættet til denne opgave består af det gennemsnitlige antal henvendelser per dag i den psykiatriske skadestue på *Virginia mental health clinic* i dagene *før*, *under* og *efter* de 12 fuldmåner fra august 1971 til juli 1972. Indkodningen af de 36 observationer i datasættet og analyse af den additive model i en 2-vejs ANOVA, hvor der tages højde både for *måned* (kategorisk variabel med 12 niveauer: *71.aug*, ..., *72.jul*) og for *fuldmåne* (kategorisk variabel med 3 niveauer: *Ja*, *Nej.efter*, *Nej.før*), fremgår af R output givet i slutningen af opgaveteksten. Besvar følgende spørgsmål:

- (a) Nedenfor ses modelvaliderings plots for den additive model på disse data. Er den additive model statistisk valid? Svaret skal begrundes.



- (b) Kan man bruge en 2-vejs ANOVA med vekselvirkning mellem *måned* og *fuldmåne* til en statistisk analyse af datasættet? Svaret skal begrundes.
- (c) Er der en signifikant sammenhæng mellem antallet af henvendelser på den psykiatriske skadestue og om det er fuldmåne? I besvarelsen af dette spørgsmål skal du opskrive en relevant nulhypotese, samt angive og fortolke den tilhørende p-værdi.
- (d) Bestem et 95% konfidensinterval for forskellen mellem antal henvendelser per dag på den psykiatriske skadestue under og før fuldmåne.
- (e) Er der signifikant forskel på antallet af henvendelser på den psykiatriske skadestue før og efter fuldmåne? Til besvarelse af dette spørgsmål skal du opskrive en relevant nulhypotese, samt beregne og fortolke den tilhørende p-værdi.

R output (lidt redigeret) til opgave 3:

```
> summary(patienter)
      måned      fuldmåne      antal
71.Aug : 3    Ja      :12   Min.   : 5.000
71.Dec : 3   Nej.efter:12  1st Qu.: 8.475
71.Nov : 3   Nej.før  :12  Median :12.850
71.Okt : 3                    Mean   :11.931
71.Sep : 3                    3rd Qu.:14.000
72.Apr : 3                    Max.   :25.000
(Other):18

> additiv.model <- lm(antal~måned+fuldmåne,data=patienter)

> plot(additiv.model)

> drop1(additiv.model,test="F")
Single term deletions

Model:
antal ~ måned + fuldmåne

Df Sum of Sq   RSS     AIC F value    Pr(>F)
<none>          127.82 73.616
måned    11    455.58 583.40 106.273  7.1285 5.076e-05 ***
fuldmåne  2     41.51 169.33  79.741  3.5726  0.04533 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```

> summary(additiv.model)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)    7.219     1.503   4.803 8.50e-05 ***
måned71.Dec   2.633     1.968   1.338 0.194555  
måned71.Nov   3.700     1.968   1.880 0.073414 .  
måned71.Okt   3.733     1.968   1.897 0.071040 .  
måned71.Sep   4.033     1.968   2.049 0.052534 .  
måned72.Apr   13.000    1.968   6.605 1.21e-06 ***
måned72.Feb   6.933     1.968   3.523 0.001916 ** 
måned72.Jan   5.033     1.968   2.557 0.017955 *  
måned72.Jul   11.033    1.968   5.606 1.23e-05 *** 
måned72.Jun   7.100     1.968   3.608 0.001563 ** 
måned72.Maj   8.600     1.968   4.370 0.000245 *** 
måned72.Mar   8.567     1.968   4.353 0.000255 *** 
fuldmåneNej.efter -1.958    0.984  -1.990 0.059149 .  
fuldmåneNej.før  -2.500    0.984  -2.541 0.018636 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 2.41 on 22 degrees of freedom
 Multiple R-squared: 0.7955, Adjusted R-squared: 0.6746
 F-statistic: 6.581 on 13 and 22 DF, p-value: 6.265e-05

Slut på opgavesættet.

Eksamensopgave i Statistisk Dataanalyse 1 (LMAB10069U)

November 2016

Alle hjælpemidler er tilladt, herunder lommeregner og computer (inkl. brug af R). Opgaverne er dog formuleret således, at der ikke er brug for computer. Der er 5 sider med 2 opgaver og i alt 12 delspørgsmål, der alle ønskes besvaret. I bedømmelsen indgår alle delspørgsmål med samme vægt. Benyttede statistiske modeller skal opskrives, og hypoteser, der testes, skal specificeres både udtrykt ved modellens parametre og i ord. Dele af opgaveteksten er givet i form af redigerede R udskrifter.

Opgave 1.

For at indsamle data til brug i undervisningen har en underviser på et statistikkursus bedt de studerende på kurset besvare et spørgeskema. De studerende skulle blandt andet angive deres køn og deres højde. I denne opgave bruges besvarelserne fra 87 kvinder og fra 35 mænd. Det antages, at disse $122 = 87 + 35$ besvarelser er repræsentative for populationerne af unge kvinder og af unge mænd. Antag at kvindernes og mændenes højder ligger i R-vektorer ved navn `kvinde` og `mand`. Betragt følgende R-output:

```
> length(kvinde)
[1] 87
> mean(kvinde)
[1] 168.3448
> sd(kvinde)^2
[1] 40.21111
> length(mand)
[1] 35
> mean(mand)
[1] 183
> sd(mand)^2
[1] 37.11765
```

Dette datasæt vil blive brugt på to forskellige måder i denne opgave. Først vil vi antage, at populationerne af højden for unge kvinder og for unge mænd hver for sig er normalfordelte med middelværdi og varians givet ved ovenstående stikprøvestørrelser beregnet i R. Besvar følgende spørgsmål:

- Hvad er sandsynligheden for at en tilfældigt udvalgt ung kvinde er højere end gennemsnitshøjden for en ung mand?
- Antag at man ikke vælger sine venner ud fra deres højde. Betragt nu et vennepar bestående af en ung kvinde og en ung mand. Hvad er sandsynligheden for at kvinden er højere end manden?

Derefter vil vi lave statistik med datasættet. Besvar følgende spørgsmål:

- (c) Hvad er den sammenvejede stikprøvespredning (på engelsk: pooled sample standard deviation) for datasættet beskrevet ved stikprøvestørrelserne ovenfor? Hvilken parameter i hvilken statistisk model er estimeret ved den sammenvejede stikprøvespredning?
- (d) Antag at variansen af højden er den samme i populationen af unge kvinder og i populationen af unge mænd. Undersøg om gennemsnitshøjden er ens for populationerne af unge kvinder og af unge mænd.
- (e) Lav et 90% konfidensinterval for forskellen mellem den gennemsnitlige højde af unge kvinder og af unge mænd.

De studerende blev også bedt om at gætte underviserens højde. Følgende tabel viser en krydstabulering mellem de studerendes køn og hvorvidt deres gæt på underviserens højde var for lavt eller for højt:

Køn af den studerende	Højdegæt var	
	for lavt	for højt
kvinde	71	16
mand	22	13

Besvar følgende spørgsmål:

- (f) Undersøg om der er en sammenhæng mellem de studerendes køn og hvorvidt de gætter for lavt eller for højt på underviserens højde. Hvis der er evidens for en sammenhæng, så beskriv denne sammenhæng.

Opgave 2.

I et forsøg undersøges effekten på plantevækst af godtning med aske. Foruden at indeholde næringsstoffer virker asken også ved at hæve pH i jorden. I 15 potter blev der tilsat aske i 3 forskellige mængder (svarende til 3, 9, og 30 t/ha), hvorefter pH i jorden og plantevæksten blev målt. Hver mængde af aske blev afprøvet i 5 potter.

Data er stillet til rådighed af Nikolaj Lunding Kindtler og Regin Rønn fra Biologisk Institut ved Københavns Universitet.

Nedenstående R output viser resultatet fra en lineær regression af plantevæksten (der hedder Y og er målt som kvadratroden af tørvægten i gram af planteskud) på pH i jorden:

```
> m1 <- lm(Y~pH,data=aske)
> summary(m1)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.71715	0.19477	8.817	7.6e-07 ***
pH	0.05010	0.03118	1.607	0.132

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'
	0.1 , , 1			

Residual standard error: 0.1703 on 13 degrees of freedom

Multiple R-squared: 0.1657, Adjusted R-squared: 0.1015

F-statistic: 2.582 on 1 and 13 DF, p-value: 0.1321

I det følgende kan det antages, at denne model er statistisk valid. Og selv om effekten af pH er ikke signifikant, skal der ikke foretages model reduktion. Besvar følgende spørgsmål:

- Opskriv den statistiske model. Angiv estimerter for alle model parametrene, og 95% konfidensintervaller for de parametre der beskriver middelværdien af Y som funktion af pH.
- Lav et 95% konfidensinterval for den gennemsnitlige plantevækst når pH er 7,0 efter godtning med aske.

Beregningshjælp: \bar{pH} og SS_{pH} kan beregnes ud fra følgende R output:

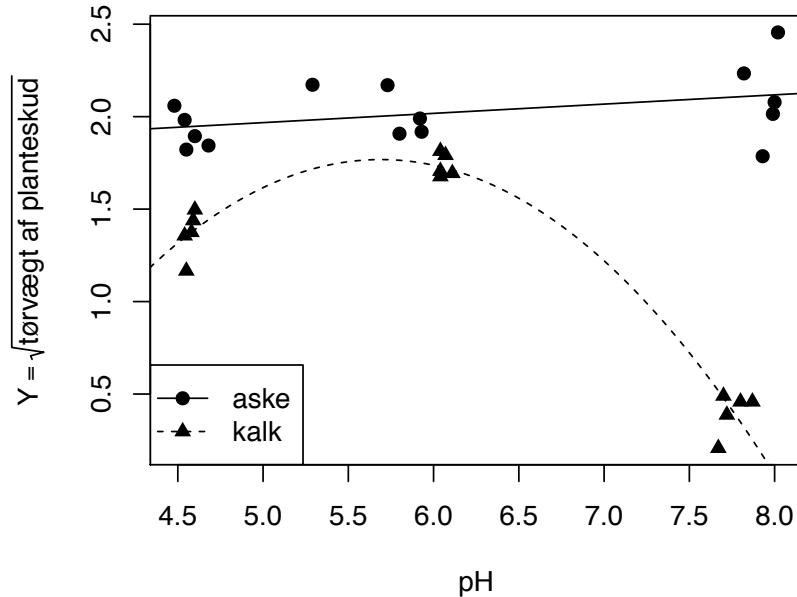
```
> mean(aske$pH)
[1] 6.085333
> sd(aske$pH)
[1] 1.460034
```

Alternativt kan SS_{pH} findes udfra formlen $SE(\hat{\beta}) = \frac{s}{\sqrt{SS_{pH}}}$.

- Lav et interval som man med 95% sandsynlighed vil forvente indeholder målingen af responsen Y fra en ny potte, der efter godtning med aske har en pH i jorden på 5,6.
- Der planlægges et nyt forsøg, hvor der bruges 4 gange så mange Potter. Antag at modellen er rigtig og hældningen i populationen er $\beta = 0,05010$. Hvilken p-værdi for effekten af pH kan man forvente i det nye forsøg?

Vink: Hvis modellen er rigtig og hældningen i populationen er $\beta = 0,05010$, så vil man forvente de samme estimerter for β og for populationsspredningen σ i det nye forsøg. Argumenter for at man vil forvente, at $SE(\hat{\beta})$ bliver halvt så stor i det nye forsøg. Brug dette til at lave et t-test for $H_0: \beta = 0$ i det nye forsøg.

Jordens pH kan også øges ved at tilsette kalk i stedet for aske. I forsøget blev det også undersøgt hvorledes plantevæksten ændrer sig med pH når der gødskes med kalk i stedet for med aske. I 15 andre potter blev der tilsat kalk i 3 forskellige mængder, som modsvarer de 3 forskellige mængder aske, hvorefter pH i jorden og plantevæksten blev målt på samme måde som før. Figuren nedenfor viser målingerne fra potterne gødsket med enten aske eller kalk, og tilhørende estimerede middelværdikurver for Y som funktion af pH.



Nedenstående R output viser resultatet fra den lineære normale model, som blev anvendt på plantevæksten som funktion af pH i jorden efter gødsning med kalk:

```
> m2 <- lm(Y~pH+I(pH^2), data=kalk)
> summary(m2)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-8.55709	1.08363	-7.897	4.29e-06 ***
pH	3.62922	0.36340	9.987	3.63e-07 ***
I(pH^2)	-0.31892	0.02931	-10.879	1.43e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1357 on 12 degrees of freedom

Multiple R-squared: 0.9548, Adjusted R-squared: 0.9473

F-statistic: 126.7 on 2 and 12 DF, p-value: 8.528e-09

I det følgende kan det antages, at denne model er statistisk valid.

Besvar følgende spørgsmål:

- (e) Opskriv den statistiske model anvendt på målingerne fra potterne gødsket med kalk. Angiv estimatorer for alle model parametrene, og 95% konfidensintervaller for de parametre der beskriver middelværdien af Y som funktion af pH.
- (f) Beregn estimatet for den pH værdi, som giver den mindste afstand mellem den gennemsnitlige plantevækst i potter gødsket med aske og med kalk. Og angiv denne mindste afstand.

Matematisk hjælp: En parabel $f(x) = a * x^2 + b * x + c$ er i sit toppunkt når $x = \frac{-b}{2*a}$. Toppunktet er der hvor $f(x)$ er mindst (hvis $a > 0$) eller størst (hvis $a < 0$).

Afslutningsvis skal det bemærkes, at det ikke er uproblematisk at gøde med aske idet asken kan indeholde tungmetaller.

Slut på opgavesættet.

Reeksamen i Statistisk Dataanalyse 1 (LMAB10069U)

Februar 2017

Alle hjælpemidler er tilladt, herunder lommeregner og computer (inkl. brug af R). Opgaverne er dog formuleret således, at der ikke er brug for computer. Der er 4 sider med 2 opgaver og i alt 12 delspørgsmål, der alle ønskes besvaret. I bedømmelsen indgår alle delspørgsmål med samme vægt. Benyttede statistiske modeller skal opskrives, og hypoteser, der testes, skal specificeres både udtrykt ved modellens parametre og i ord. Dele af opgaveteksten er givet i form af redigerede R udskrifter.

Opgave 1.

Google afprøver løbende nye søgealgoritmer på deres hjemmeside www.google.com. En *succesfuld søgning* er en søgning, hvor brugeren klikker på et af de foreslænde links. Hvis man som bruger ikke er tilfreds med de foreslæde links, så vil man typisk lave en ny søgning på relaterede ord. Dette betegnes som en *ikke-succesfuld søgning*. I et forsøg undersøgte Google resultaterne fra 10000 søger, hvoraf 5000 blev lavet med den *nuværende* søgealgoritme og 2500 blev lavet med hver af to nye algoritmer (som kaldes for *A* og *B* nedenfor). For ikke at influere på deres opførelse fik de berørte brugere på www.google.com ikke at vide, at de deltog i et forsøg. Og det er jo nok de færreste brugere af www.google.com der tænker over, at søgeresultaterne kan være genereret af forskellige matematiske algoritmer.

Resultaterne fra forsøget er opsummeret i 2×3 -tabellen nedenfor:

Søgealgoritme:	<i>nuværende</i>	<i>A</i>	<i>B</i>	Total
<i>succesfuld søgning</i>	3511	1749	1818	7078
<i>ikke-succesfuld søgning</i>	1489	751	682	2922
Total	5000	2500	2500	10000

Besvar følgende spørgsmål:

- Angiv det forventede antal *succesfulde* og *ikke-succesfulde søger* for hver af de tre søgealgoritmer under nulhypotesen, at søgealgoritmerne har samme sandsynlighed for en *succesfuld søgning*.
- Har de tre søgealgoritmer samme sandsynlighed for at give en *succesfuld søgning*? Svaret skal begrundes.
- Angiv et 95% konfidensinterval for sandsynligheden for en *succesfuld søgning* for hver af de tre søgealgoritmer. Hvilken søgealgoritme er bedst?
- Antag, at Google besluttede de i en overgangsperiode ville bruge søgealgoritme *B* i 60% af søgerne på www.google.com, og den *nuværende* søgealgoritme i de resterende søger. Hvor stor en andel af søgerne i overgangsperioden ville man forvente var *succesfulde*?

- (e) Betragt en tilfældig udvalgt bruger blandt de 10000 brugere der deltog i forsøget. Beregn følgende to sandsynligheder:
- (1) Sandsynligheden for at brugeren foretog en *successful* *søgning* med søgealgoritme B .
 - (2) Den betingede sandsynlighed (på engelsk: conditional probability) for at brugeren brugte søgealgoritme B givet at søgningen var *successful*.

Opgave 2.

For at indsamle data til brug i undervisningen har en underviser på et statistikkursus bedt de studerende på kurset besvare et spørgeskema. De studerende blev blandt andet bedt om at angive deres køn og deres højde, og om at gætte på underviserens højde. Antag, at observationerne af disse tre størrelser findes i variablerne `sex`, `height` og `height.guess` inde i en dataframe ved navn `sd1`. Her er `sex` en kategorisk variabel (Mand eller Kvinder), mens `height` og `height.guess` er kontinuerte variable der angiver højde målt i cm. Betragt følgende R output:

```
> m1 <- lm(height.guess~sex*height,data=sd1)
> summary(m1)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 166.94594   12.39836 13.465   <2e-16 ***
sexMand     -11.95713   25.45080 -0.470    0.639    
height       0.08205   0.07360  1.115    0.267    
sexMand:height 0.06471   0.14202  0.456    0.650    

```

Residual standard error: 4.328 on 118 degrees of freedom
Multiple R-squared: 0.03459, Adjusted R-squared: 0.01005
F-statistic: 1.409 on 3 and 118 DF, p-value: 0.2435

```
> m2 <- lm(height.guess~sex+height,data=sd1)
> summary(m2)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 164.02078   10.57116 15.516   <2e-16 ***
sexMand     -0.37579    1.25603 -0.299    0.765    
height       0.09943   0.06273  1.585    0.116    

```

Residual standard error: 4.314 on 119 degrees of freedom
Multiple R-squared: 0.03289, Adjusted R-squared: 0.01664
F-statistic: 2.024 on 2 and 119 DF, p-value: 0.1367

```

> m3 <- lm(height.guess~height,data=sd1)
> summary(m3)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 166.26466   7.42161 22.403 <2e-16 ***
height       0.08579   0.04296  1.997  0.0481 *

```

Residual standard error: 4.297 on 120 degrees of freedom
 Multiple R-squared: 0.03217, Adjusted R-squared: 0.0241
 F-statistic: 3.988 on 1 and 120 DF, p-value: 0.04808

Besvar følgende spørgsmål:

- (a) Opskriv de statistiske modeller svarende til R-modellerne m1, m2 og m3.
- (b) Lav modelreduktion startende med modellen m1. Angiv slutmodellen samt p-værdierne for de hypotese tests der laves undervejs.

R-modellen m3 er en lineær regression af height.guess på height. I resten af opgaven bruges denne statistiske model. Besvar følgende spørgsmål:

- (c) Angiv estimerter for alle model parametrene, og 95% konfidensintervaller for de parametre der beskriver middelværdien af højdegættet som funktion af de studerendes højde.
- (d) Lav et 95% konfidensinterval for det gennemsnitlige gæt på underviserens højde af studerende der selv er 168 cm høje.

Beregningshjælp: \bar{height} og SS_{height} kan beregnes ud fra følgende R output:

```

> mean(sd1$height)
[1] 172.5164
> sd(sd1$height)
[1] 9.093167

```

Alternativt kan SS_{height} findes udfra formlen $SE(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{SS_{height}}}$, hvor $\hat{\beta}$ og $\hat{\sigma}$ er estimerterne henholdsvis for hældningen mht. height og for residual spredningen i modellen m3.

- (e) En af de studerende var ikke til forelæsningen den dag data blev indsamlet. Antag, at denne studerende er 183 cm høj. Lav et interval som man med 95% sandsynlighed vil forvente indeholder målingen af gættet på underviserens højde fra den pågældende studerende.
- (f) Antag, at gennemsnithøjden af de mandlige og kvindelige studerende på kurset er henholdsvis 183 og 168 cm. Lav et 95% konfidensinterval for den gennemsnitlige forskel på mændenes og kvindernes gæt på underviserens højde.

De statistiske beregning lavet ovenfor er kun pålidelige hvis de tilhørende modelantagelser kan valideres.

- (g) Opskriv nogle linier med R kode der kan bruges til at lave grafisk modelvalidering for modellen m3.

Slut på opgavesættet.

Statistisk Dataanalyse 1 (StatDat1, LMAB10069U)

Eksamens, november 2017

Fire timers skriftlig prøve. Alle hjælpemidler er tilladt, herunder lommeregner og computer (fx brug af R), og besvarelsen må gerne skrives med blyant. Du kan *ikke aflevere elektronisk*, heller ikke på vedlagte USB-stick.

Der er 4 opgaver med i alt 13 delspørsgsmål. Alle delspørsgsmål indgår med samme vægt i bedømmelsen. Husk at de fleste spørsgsmål kan besvares uafhængigt af hinanden.

Data til opgave 1 og opgave 3 udleveres på en USB-stick. Filnavnene fremgår af opgaveteksten. USB-sticken skal afleveres efter eksamen, men udelukkende for at den kan genbruges. Den kan ikke indgå som en del af besvarelsen. Der er R-kode og R-output til opgave 2.

Opgave 1

Det bliver med jævne mellemrum diskuteret i USA om den liberale våbenlovgivning har negative konsekvenser, fx om den lette adgang til våben fører til flere selvmord. For at undersøge dette har man indsamlet oplysninger fra hver af de 50 amerikanske stater. Data er tilgængelige på den vedlagte USB-stick som guns.txt og guns.xlsx. Der er en linie per stat og følgende variable:

- State: Navnet på staten
- GunOwnerPct: Udbredelsen af skydevåben, målt som procentdelen af husholdninger der ejer mindst et skydevåben
- SuicideRate: Antal selvmord per 100000 indbyggere
- Law: Har værdien Yes eller No afhængig af om staten har love (mindst en) der lægger restriktioner på våbensalg uddover det der gælder i hele USA

I de første to spørsgsmål skal du kun bruge variablene GunOwnerPct og SuicideRate.

1. Lav en figur der illustrerer sammenhængen mellem udbredelsen af skydevåben og selvmordsraten. Der skal være en skitse af figuren i besvarelsen.

Angiv på baggrund af figuren en statistisk model der gør det muligt at estimere sammenhængen, og angiv estimerater for samtlige parametre i modellen.

2. Brug modellen til at undersøge om der er sammenhæng mellem våbenudbredelse og selvmordsrate.

Bestem et estimat og et 95% konfidensinterval for den forventede selvmordsrate for en fiktiv stat med en våbenudbredelse på 45%.

Man må formode at love der sætter begrænsninger for våbensalget, begrænser udbredelsen af våben. Det er derfor fornuftigt at inddrage variablen Law i analysen, og de sidste spørsgsmål skal derfor besvares på baggrund af følgende modelfit, hvor guns er navnet på det indlæste datasæt:

```
lm(SuicideRate ~ GunOwnerPct + Law, data=guns)
```

Gennemsnittet af GunOwnerPct er 27.15% for stater der har mindst en restriktiv lov (Law=Yes) og 45.19% for stater der ikke har en restriktiv love (Law>No).

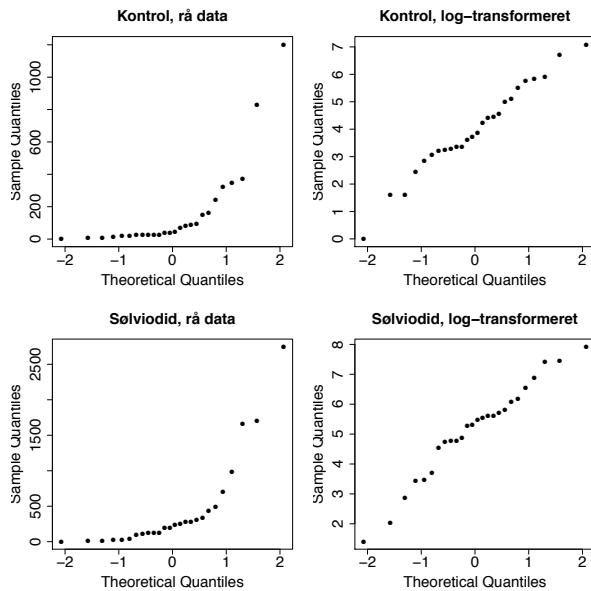
3. Bestem estimerer for den forventede selvmordsrate for to fiktive stater: en stat med mindst en restriktiv lov og en våbenudbredelse på 27.15%; og en stat uden restriktive love og en våbenudbredelse på 45.19%.
4. Tyder data på at der er en effekt af restriktive love på selvmordsraten, når man tager højde for våbenudbredelsen?

Opgave 2

Data til denne opgave består af regnmængder fra 52 skyer. Halvdelen af skyerne blev behandlet med sølviodid i eksperimentet, mens den anden halvdel var kontrolskyer som ikke blev behandlet. Regnmængden er målt i *acre-feet*, som er den mængde vand der kræves for at dække et areal på 1 acre (4047 m^2) i en højde på 1 fod (0.305 m).

Data er indlæst i datasættet `regnData` i R med to variable: `behandling` der enten er `kontrol` eller `sølvIodid`, og `regn` der angiver den observerede regnmængde.

Figuren nedenfor viser fire QQ-plots: De øverste plots er for kontrolskyerne, de nederste er for skyerne behandlet med sølviodid. Til venstre er QQ-plots lavet for de rå (ikke-transformerede) data, til højre for log-transformerede værdier.



Første spørgsmål vedrører kun de 26 kontrolskyer.

1. Forklar kortfattet hvorfor det er mere fornuftigt at analysere de log-transformerede værdier fra kontrolskyerne end de ikke-transformerede værdier som en normalfordelt stikprøve.
Bestem et 95% konfidensinterval for middelværdien af log-transformeret regnmængde for kontrolskyer. Du kan benytte at gennemsnittet for de 26 værdier af `log(regn)` fra kontrolskyer er 3.990, og at stikprøvespredningen er $s = 1.642$.

Vi skal nu interesser os for effekten af sølviodidbehandlingen.

Der er R-kode og R-output sidst i opgaven som kan benyttes ved besvarelserne. Dele af outputtet er erstattet af XXXX. Det er med vilje og værdierne kan beregnes udfra den givne information og en smule R-kode.

2. Angiv en statistisk model der kan bruges til at sammenligne regnmængden for kontrolskyer og skyer behandlet med sølviodid. Udfør et hypotesetest der belyser om sølviodidbehandlingen har en effekt på regnmængden.
3. Angiv et estimat og et 95% konfidensinterval for forskellen mellem de forventede værdier af logaritmen til regnmængden for de to grupper af skyer.
Angiv derefter et estimat og et 95% konfidensinterval for den procentvise forøgelse af regnmængden når skyer tilføres sølviodid.

Uddrag af R-kørsel. Dele af outputtet er erstattet af XXXX. Det er med vilje og værdierne kan beregnes udfra den givne information og en smule R-kode.

```
> model <- lm(log(regn) ~ behandling, data=regnData)
> summary(model)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.9904     0.3179    XXXX    XXXX
behandlingsolvIodid 1.1438     0.4495    XXXX    XXXX
---
Residual standard error: 1.621 on 50 degrees of freedom

> confint(model)
                2.5 %   97.5 %
(Intercept)      3.351948 4.628864
behandlingsolvIodid 0.240865 2.046697
```

Opgave 3

Data til denne opgave stammer fra et eksperiment, hvor rodlængden blev målt for 40 planter 10 dage efter såning. Planterne blev dyrket enkeltvis i potter, som stod to forskellige steder (sted A og B, 20 planter per sted). Jorden i potterne var præpareret med gødning, men på fire forskellige måder (se nedenfor). Data er tilgængelige på den vedlagte USB-stick som `rodlaengde.txt` og `rodlaengde.xlsx`. Der er en linie per plante og følgende variable:

- `dosis`: Behandlingsvariabel. Har værdierne `lav` svarende til lav dosis, `mellem1` svarende til mellemstør dosis givet som en enkelt behandling, `mellem2` svarende til mellemstør dosis givet som to behandlinger, `høj` svarende til høj dosis.
- `sted`: Stedet hvor planten er dyrket. Har værdierne A og B.
- `lgd`: Den målte rodlængde i cm.

- Fit modellen for tosidet variansanalyse (tosidet ANOVA) *med vekselvirkning* med rodlængden som responsvariabel, og udfør modelkontrol. Besvarelsen skal bestå af en linie R-kode med `lm`-kommandoen, skitser af de relevante figurer og kommentarer til figurerne.
- Undersøg om effekten af dosis på rodlængden er forskellig på sted A og sted B.

I de næste spørgsmål skal du benytte modellen for tosidet ANOVA *uden vekselvirkning* med rodlængden som responsvariabel—uanset hvad du har svaret i spørgsmål 1 og 2.

- Bestem estimatet for forventet rodlængde for en plante fra sted B, som har fået mellemstor dosis gødning givet som to behandlinger (`mellem2`).

For hvilken af de otte kombinationer af dosis og sted er den forventede rodlængde størst? Svaret skal naturligvis begrundes.

- Angiv estimat og 95% konfidensinterval for *forskellen* i forventet rodlængde mellem planter der har fået høj dosis gødning og planter der har fået lav dosis gødning.

Undersøg med et hypotesetest om der er forskel mellem forventet rodlængde på sted A og sted B.

Opgave 4

En ingrediens i kosmetikprodukter mistænkes for at øge risikoen for en ellers sjælden hudsygdom. For at undersøge sammenhængen, har man udvalgt 223 kvinder med sygdommen (cases) og desuden 446 kvinder uden sygdommen (controls). Ved udvælgelsen ved man ikke om kvinderne har været eksponeret for ingrediensen, men det kan afgøres med en blodprøve.

Resultatet fremgår af tabellen nedenfor.

	Eksponeret	Ikke eksponeret	Total
Har sygdommen	54	169	223
Har ikke sygdommen	76	370	446

- Undersøg med et hypotesetest om sandsynligheden for at en kvinde har været eksponeret, afhænger af om hun har sygdommen eller ej. Hvad er konklusionen i forhold til sammenhængen mellem ingrediensen og sygdommen?

Studiet er konstrueret således at en tredjedel af de undersøgte kvinder har sygdommen, men forekomsten af sygdommen i befolkningen er kun 0.5%. Hvis man udtager en kvinde tilfældigt fra befolkningen, er sandsynligheden altså 0.5% for at vælge en der har sygdommen. Antag desuden at sandsynlighederne for at kvinder med og uden sygdommen har været eksponeret til ingrediensen, er 0.242 og 0.170. Dette svarer til tabellen ovenfor.

- Bestem sandsynligheden for at en tilfældig kvinde har været eksponeret. Bestem derefter den betingede sandsynlighed for at en kvinde der har været eksponeret, har sygdommen.

Statistisk Dataanalyse 1 (StatDat1, LMAB10069U)

Eksamens, januar 2018

Fire timers skriftlig prøve. Alle hjælpemidler er tilladt, herunder lommeregner og computer (fx brug af R), og besvarelsen må gerne skrives med blyant. Du kan *ikke aflevere elektronisk*, heller ikke på vedlagte USB-stick.

Der er 3 opgaver med i alt 13 delspørsgsmål. Alle delspørsgsmål indgår med samme vægt i bedømmelsen. Husk at de fleste spørsgsmål kan besvares uafhængigt af hinanden. Alle svar skal begrundes. Data til opgave 1 og opgave 2 udleveres på en USB-stick. Filnavnene fremgår af opgaveteksten. USB-sticken skal afleveres efter eksamen, men kun for at den kan genbruges.

Opgave 1

På kurset *Sandsynlighedsregning og Statistik* i 2017/18 bad underviseren de studerende om at gætte på antallet af punkter i tre figurer. Denne opgave handler om den ene figur (figur 2) og der er gæt fra 182 studerende. Data er tilgængelige på den vedlagte USB-stick som `ss2017-18.txt` og `ss2017-18.xlsx`. Der er en linie per studerende og følgende variable.

- **studie:** Det studie som den studerende er indskrevet på. De mulige værdier er **Matematik**, **Mat0k** (matematik-økonomi) og **Aktuar** (aktuar/forsikringsmatematik)
 - **kon:** Den studerendes køn, enten **Mand** eller **Kvinde**
 - **figur2:** Den studerendes gæt på antal punkter i figuren
1. Forklar kortfattet hvorfor det er naturligt at benytte en tosidet variansanalyse til disse data. Angiv R-kode der kan bruges til at estimere følgende to modeller, og angiv residualspreddingen (Residual standard error) for begge modeller:
 - En tosidet variansanalyse *med vekselvirkning* hvor du bruger variablen `figur2` som responsvariabel og de andre variable som forklarende variable.
 - En tosidet variansanalyse *med vekselvirkning* hvor du bruger variablen `log(figur2)` som responsvariabel og de andre variable som forklarende variable. Husk til senere brug at `log` er den naturlige logaritme.
 2. Udfør modelkontrol for hver af de to modeller fra spørsgsmål 1. Besvarelsen skal bestå af skitser af de relevante figurer og kommentarer til figurerne, herunder argumenter for at modellen med `log(figur2)` som responsvariabel er at foretrække.
 3. Undersøg med et hypotesetest om der er vekselvirkning mellem køn og studie, og forklar kortfattet hvad resultatet betyder. Du skal benytte `log(figur2)` som responsvariabel.

I de næste spørsgsmål skal du benytte modellen for tosidet ANOVA *uden vekselvirkning* uanset hvad du har svaret i spørsgsmål 3. Du skal benytte `log(figur2)` som responsvariabel.

4. Angiv et estimat og et 95% konfidensinterval for den forventede værdi af `log(figur2)` for kvindelige aktuarstuderende.

Det sande antal punkter i figuren er 142. Tyder data på at kvindelige aktuarstuderende (som population) gætter for højt, for lavt, eller ingen af delene? Svaret skal begrundes, og du kan benytte at $\log(142) = 4.956$.

5. Angiv et estimat for forskellen mellem kvinder og mænd i forventet værdi af $\log(\text{figur}2)$.
Angiv derefter et estimat for den faktor som kvinders gæt er højere end mænds gæt. Er der signifikant forskel på mænd og kvinder?
6. Undersøg med *et enkelt* hypotesetest om studerende fra de tre forskellige studier (som populationer) gætter forskelligt på antallet af punkter i figuren.

Opgave 2

Data til denne opgave består af kropsmålinger fra 243 mænd. For hver mand har man blandt andet målt omkredsen ved hofte og mave, begge dele i cm. Desuden har man bestemt mændenes fedtprocent med en præcis målemetode baseret på opdriften ved undervandsvejning. Man er interesseret i at kunne prædiktere fedtprocenten ved hjælp af hofte- og/eller maveomkreds.

Data er tilgængelige i filerne johnson-fatpct.txt og johnson-fatpct.xlsx på den vedlagte USB-stick. Der er en linie per person og følgende variable:

- bodyfat: Fedtprocent
- hip: Omkreds ved hofte, målt i cm
- abdomen: Omkreds ved mave, målt i cm

Du skal i hele opgaven bruge variablen **bodyfat** som responsvariabel.

1. Lav en figur der illustrerer sammenhængen mellem maveomkreds og fedtprocent. Der skal være en skitse af figuren i besvarelsen.
Angiv på baggrund af figuren en statistisk model der gør det muligt at estimere sammenhængen.
2. Angiv estimerater for samtlige parametre i modellen fra spørgsmål 1.
Betrægt to mænd der har maveomkreds på henholdsvis 100 cm og 110 cm. Bestem et estimat for forskellen i forventet fedtprocent mellem de to mænd.
3. Fit den lineære regressionsmodel hvor du bruger hofteomkredsen som den eneste forklarende variabel, og angiv estimatet for regressionskoefficienten hørende til hofteomkreds.
Fit derefter den multiple regressionsmodel hvor du inddrager både maveomkreds og hofteomkreds som forklarende variable, og angiv estimatet for regressionskoefficienten hørende til hofteomkreds.
Forklar kortfattet hvad forskellen på de to angivne estimerater kan skyldes.
4. I dette spørgsmål skal du bruge den multiple lineære regression fra spørgsmål 3. Bestem et 95% konfidensinterval og et 95% prædiktionsinterval for en mand med maveomkreds på 85 cm og hofteomkreds på 98 cm.
Er det usædvanligt for en mand med maveomkreds på 85 cm og hofteomkreds på 98 cm at have en fedtprocent på 17? Svaret skal begrundes.

Opgave 3

Forskere i New England har udvalgt 73 sjældne plantearter tilfældigt og vurderet deres udbredelse med fem års mellemrum, nemlig i 2012 og 2017. Den samme metode og den samme skala er brugt begge år, og alle arter er vurderet begge år.

Forskellen i udbredelse mellem 2012 og 2017 er beregnet for hver af de 73 arter som

$$\text{forskel} = \text{udbredelse i 2017} - \text{udbredelse i 2012}$$

Forskellen er indlæst i R som variablen `forskel` og du kan benytte følgende værdier:

```
> mean(forskel)
[1] 3.353576
> sd(forskel)
[1] 8.303946
```

1. Forklar kortfattet hvorfor data fra de to år skal opfattes som to parrede stikprøver snarede end som to uparrede (uafhængige) stikprøver.

Angiv et estimat og et 95% konfidensinterval for den forventede forskel i udbredelse mellem 2012 og 2017.

Man vil gerne undersøge om fredning er et effektivt redskab til at forbedre sjældne planters udbredelse. For hver af de 73 arter har man derfor koblet deres fredningsstatus med ændringen i udbredelse fra 2012 til 2017. Man har valgt ikke at bruge de specifikke værdier i variablen `forskel`, men udelukkende fortegnet, dvs. om artens udbredelse er vokset eller aftaget.

Data fremgår af tabellen nedenfor, og i resten af opgaven skal du kun bruge tallene fra tabellen.

	Ikke fredet	Fredet
Udbredelse aftaget	18	8
Udbredelse vokset	15	32

2. Undersøg med et hypotesetest om der er sammenhæng mellem fredningsstatus og ændring i udbredelse.
3. Angiv et estimat for den betingede sandsynlighed for at en plantearts udbredelse vokser givet at den fredet, samt den betingede sandsynlighed for at en plantearts udbredelse vokser givet at den ikke er fredet.