Eksamen i Statistisk Dataanalyse 2
(LMAF10070)

3. april 2014

Alle sædvanlige hjælpemidler, herunder bøger, noter, `R`-programmer og lommeregner samt brug af programmet `R` på egen PC, er tilladt. Det er *ikke* tilladt at benytte PC til nogle former for aktivitet, som involverer opkobling til et netværk eller kommunikation med andre. Det er tilladt at skrive med blyant. Opgavesættet består af 12 sider med i alt 3 opgaver, der indgår med vægtningen 45 %, 25 % og 30 % i bedømmelsen.

Til besvarelse af opgave 1 har du fået udleveret en USB-nøgle med et datasæt, som du skal indlæse og anvende i `R` på din egen PC for at kunne besvare opgaven. Til opgave 2 er der vedlagt udvalgte `R`-udskrifter, som kan benyttes i besvarelsen (det er ikke sikkert at alle dele af udskriften skal benyttes). Husk at det er vigtigt at specificere de statistiske modeller og hypoteser du bruger, og at komme med konklusioner på analyserne.

Opgave 1 (5 spørgsmål)

Ved Institut for Produktionsdyr og Heste på Københavns Universitet har man over en årrække udviklet en målemetode til vurdering af halthed hos heste. Konkret måler man accelerationen af hesten mest den traver, og på baggrund heraf udregnes en score `S`, der måler graden af symmetri i hestens bevægelsesmønster. Høje værdier af `S` svarer til den største grad af symmetri.

Af praktiske årsager er det hensigtsmæssigt at foretage målingerne mens hestene traver i cirkler omkring en person, som står stille i centrum af cirklen og holder hesten i cirkelbevægelsen ved brug af et reb. Vi betragter i denne opgave et datasæt, hvor man har ladet 27 raske (dvs. ikke-halte) heste (`hest`) trave i cirkler af varierende diameter (`diam=8,10,12,14,16` meter). For hver diameter har man desuden ladet hesten løbe rundt både mod højre og mod venstre (`side` med niveauerne `H` og `V`). Data til opgaven er venligst stillet til rådighed af Maj Halling Thomsen.

Data er udleveret på vedlagte USB-nøgle under filnavnet `data1.txt` og for at besvare opgaven fuldstændigt, vil det være nødvendigt at køre udvalgte `R`-kommandoer på din egen medbragte computer. Data kan f.eks. indlæses ved brug af kommandoen

```
data1<-read.table(file.choose(),header=T)
```

hvor du vælger filen `data1.txt`. De første linjer i datasættet er organiseret som vist nedenfor.

```
> head(data1,12)
```

```
   hest side diam        S
1   G01    H    8 4.723692
2   G01    H   10 4.927638
3   G01    H   12 5.126423
4   G01    H   14 5.475348
5   G01    H   16 5.000378
6   G01    V    8 4.297942
7   G01    V   10 4.290536
8   G01    V   12 4.251726
9   G01    V   14 4.459603
10  G01    V   16 4.589719
11  G02    H    8 4.430027
12  G02    H   10 4.578747
```

Du kan i hele opgaven se bort fra vekselvirkninger mellem `hest` og `diameter`, men øvrige vekselvirkninger ønskes medtaget ved de statistiske analyser. Du skal benytte `S` som responsvariabel i dine statistiske modeller, og det er *ikke* en del af opgaven, at du skal bruge tid på modelkontrol.

Ved besvarelse af delspørgsmål 1.-2. nedenfor skal du opfatte variablen `diam` som en faktor med 5 niveauer.

1. Opskriv en statistisk model der kan tages som udgangpunkt for en analyse af, hvordan symmetriscoren (`S`) afhænger af omløbsretning (`side`) samt af cirklens diameter (`diam`). Tegn et faktordiagram for modellen.

2. Foretag modelreduktion i modellen fra spørgsmål 1. og angiv parameterestimater samt 95 %-konfidensintervaller for middelværdi- og variansparametre i slutmodellen.

I det følgende fortsættes analysen ovenfor med henblik på at undersøge muligheden for at lade diameteren (`diam`) indgå som en numerisk variabel.

3. Tag udgangspunkt i din slutmodel fra delspørgsmål 2. Foretag et statistisk test af, om det er rimeligt at antage, at symmetriscoren (`S`) afhænger lineært af diameteren. Du bedes tydeligt opskrive de modeller, du tester imod hinanden.

4. Et tidligere studie har vist, at den forventede værdi af symmetriscoren `S` for en rask hest (=ikke-halt) som løber ligeud er 5.63. Benyt analysen fra delspørgsmål 1.-3. til at diskutere, hvordan cirklens diameter influerer på symmetrien i hestens gang.

5. Brug resultaterne fra delspørgsmål 1.-3. og om nødvendigt nogle ekstra analyser til at diskutere, om datasættet underbygger en påstand om, at heste har en foretrukken omløbsretning, når de løber i cirkler.

Opgave 2 (3 spørgsmål)

I forbindelse med udviklingen af et nyt mejeriprodukt (yoghurt) ønsker man at eksperimentere med sukkerindhold, aroma og konsistens. Konkret har man besluttet sig for at lave et eksperiment, hvor der afprøves to forskellige niveauer af hver af faktorerne sukker (`s` med værdierne `1=lav`, `2=høj`), aroma (`a` med værdierne `1=ikke tilsat` ,`2=tilsat`) og konsistens (`k` med værdierne `1=vandig`, `2=tyk`). For at afprøve produkterne har man tænkt sig at invitere nogle smagsdommere (-givet ved faktoren `d`) til hver at afprøve 4 forskellige af de i alt 8 mulige kombinationer af de tre faktorer. Smagsdommerne giver hvert produkt en score (`=y`) på en kontinuert (dvs numerisk) skala, f.eks. fra 0 til 100.

I første omgang planlægges det at invitere 8 smagsdommere. Af praktiske årsager beslutter man, at hver smagsdommer enten afprøver produkter med `vandig` konsistens eller produkter med `tyk` konsistens, således at man benytter følgende forsøgsplan.

`kombination nr`	`s`	`a`	`k`	d=1	d=2	d=3	d=4	d=5	d=6	d=7	d=8
1	1	1	1					x	x	x	x
2	1	1	2	x	x	x	x				
3	1	2	1					x	x	x	x
4	1	2	2	x	x	x	x				
5	2	1	1					x	x	x	x
6	2	1	2	x	x	x	x				
7	2	2	1					x	x	x	x
8	2	2	2	x	x	x	x				

1. Opskriv et faktordiagram og en tilhørende statistisk model til analyse af forsøget. Det er tilladt at se bort fra vekselvirkninger med smagsdommer (`d`). Hvilken type forsøg er der tale om, og hvordan bør randomiseringen foretages?

Man beslutter sig for i stedet at udføre forsøget således, at man for det første par (`1-2`) af smagsdommere konfunderer 3-faktorvekselvirkningen (`s` × `a` × `k`) med person, mens man for hvert af de øvrige 3 par af smagsdommere (`3-4`,`5-6` og `7-8`) konfunderer en af de 3 parvise vekselvirkninger (`s` × `a`, `s` × `k` og `a` × `k`) med person.

2. Opskriv den tilhørende forsøgsplan i et skema som det, der er vist ovenfor.

Det er på ret forhånd klart, at ingen vil være interesseret i en vandig yoghurt, uden sukker og uden det tilsatte aromastof (-svarende til `kombination nr 1` i skemaet ovenfor). I praksis er der derfor kun 7 forskellige yoghurt kombinationer (`2-8`), som skal afprøves i forsøget.

3. Lav en forsøgsplan som viser, at det er muligt at lave et balanceret ufuldstændigt blokforsøg med 7 smagsdommere, som hver smager på 4 af de 7 yoghurtkombinationer. Du kan med fordel tage udgangspunkt i den forsøgsplan, som er påbegyndt nedenfor, hvor du blot skal angive, hvilke yoghurt kombinationer, der skal afprøves af smagsdommer `d=6` og `d=7`.

kombination nr	s	a	k	d=1	d=2	d=3	d=4	d=5	d=6	d=7
2	1	1	2		x	x				
3	1	2	1		x	x	x	x		
4	1	2	2	x		x		x		
5	2	1	1	x		x	x			
6	2	1	2	x	x			x		
7	2	2	1	x	x		x			
8	2	2	2				x	x		

Opgave 3 (3 spørgsmål)

Som en del af et større projekt har man interesseret sig for søvnkvaliteten hos patienter med leddegigt. På baggrund af et spørgeskema har man for hver patient udregnet en score (kaldet `psqi`) fra `1-21`, hvor lave værdier svarer til god søvnkvalitet. Det er desuden velkendt at både smerter og depression kan hænge sammen med søvnkvalitet, ligesom der kan være forskelle mellem mænd og kvinders søvnkvalitet. Det samlede datasæt indholder derfor variablen `gender` med værdierne `Kvinde` og `Mand`, variablen `depression` med værdierne `1` (=depression) og `0` (=ikke depression) samt variablen `vas.smerter`, der måler patientens almene smerteniveau på en såkaldt VAS-skala fra `0-100` (-hvor 100 svarer til værst tænkelige smerte!).

Data er venligst stillet til rådighed af Katrine Bjerre Løppenthin. Et udpluk af datasættet kan ses nedenfor.

```
> head(data3,10)

   psqi gender depression vas.smerter
1     7   Mand          0           3
2    10 Kvinde          0          40
3     9 Kvinde          0          55
4     5   Mand          0           3
5     2   Mand          0          25
6    11 Kvinde          0          50
7     7 Kvinde          0          51
10    1   Mand          0           1
11   13 Kvinde          0          84
12   19 Kvinde          0          43
```

Ved besvarelsen af følgende 3 delspørgsmål skal du benytte (dele af) `R`-udskriften nedenfor. For en god ordens skyld gøres der opmærksom på, at `sqrt(psqi)` er det samme som $\sqrt{\texttt{psqi}}$.

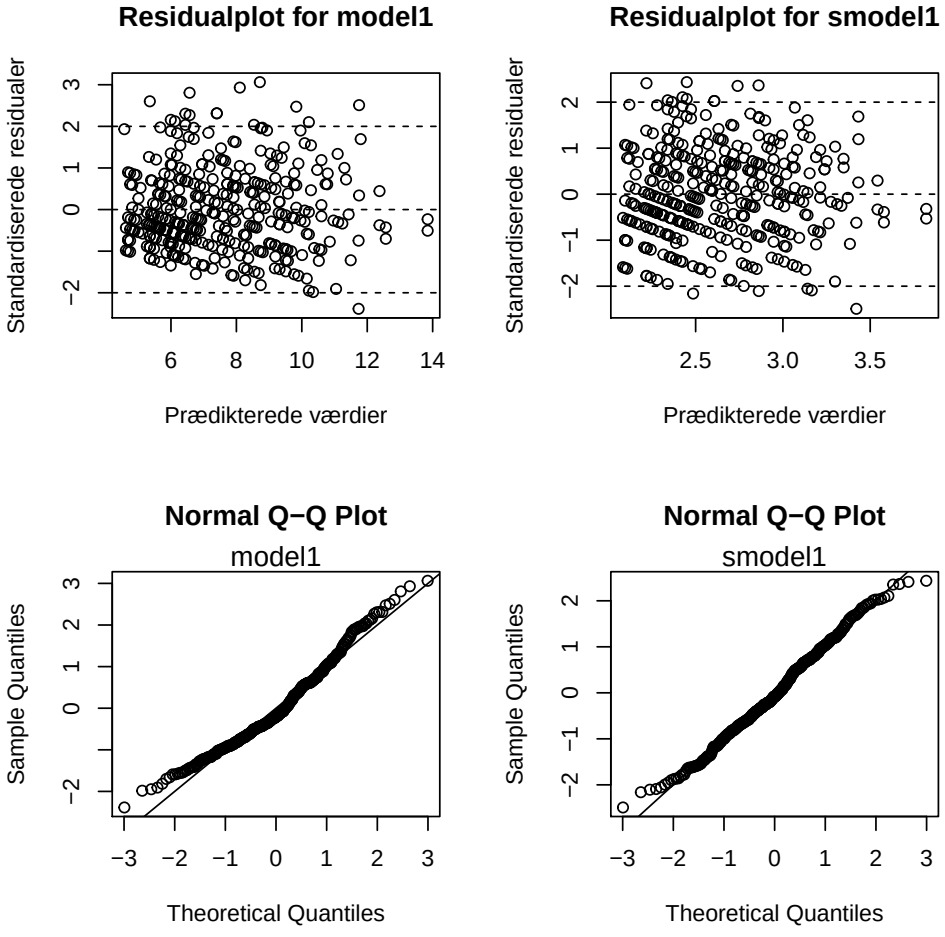
1. Benyt `R`-udskriften til at foreslå en statistisk model, der med rimelighed kan benyttes som udgangspunkt for en analyse af, hvordan søvnkvalitet hænger sammen med de øvrige variable i datasættet. Husk at begrunde dit valg af model.

2. Foretag en reduktion af modellen fra delspørgsmål 1. med henblik på at undersøge, hvilke variable der er associeret med (dvs hænger sammen med) søvnkvalitet.

Undervejs, skal du tydeligt gøre rede for, hvilke modeller du tester imod hinanden, ligesom du bedes angive teststørrelser og p-værdier svarende til de enkelte test, som du foretager. Angiv alle parameterestimater i slutmodellen, samt 95 %-konfidensintervaller for parametrene, som indgår i beskrivelsen af middelværdistrukturen. Skriv også en konklusion i ord om, hvad slutmodellen udtrykker.

3. Benyt din slutmodel fra delspørgsmål 2. til at angive et estimat og et 95 %-konfidensinterval for søvnkvaliteten for depressive mænd med `vas.smerter=30`.

```
> ### Nogle statistiske modeller og figurer:
> data3$depfac<-factor(data3$depression)
> model1<-lm(psqi~gender*depfac+vas.smerter,data3)
> model2<-lm(psqi~gender+depfac+vas.smerter,data3)
> model3<-lm(psqi~gender+vas.smerter,data3)
> model4<-lm(psqi~depfac+vas.smerter,data3)
> model5<-lm(psqi~vas.smerter,data3)
> smodel1<-lm(sqrt(psqi)~gender*depfac+vas.smerter,data3)
> smodel2<-lm(sqrt(psqi)~gender+depfac+vas.smerter,data3)
> smodel3<-lm(sqrt(psqi)~gender+vas.smerter,data3)
> smodel4<-lm(sqrt(psqi)~depfac+vas.smerter,data3)
> smodel5<-lm(sqrt(psqi)~vas.smerter,data3)
```



```
> ### Nogle statistiske test:
> anova(model2,model1)

Analysis of Variance Table

Model 1: psqi ~ gender + depfac + vas.smerter
Model 2: psqi ~ gender * depfac + vas.smerter
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1    359 5000.1
2    358 4966.1  1    34.013 2.452 0.1183

> anova(model3,model2)

Analysis of Variance Table

Model 1: psqi ~ gender + vas.smerter
Model 2: psqi ~ gender + depfac + vas.smerter
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    360 5097.4
2    359 5000.1  1    97.224 6.9805 0.008601 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(model4,model2)

Analysis of Variance Table

Model 1: psqi ~ depfac + vas.smerter
Model 2: psqi ~ gender + depfac + vas.smerter
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    360 5051.0
2    359 5000.1  1    50.899 3.6544 0.05672 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(model5,model3)

Analysis of Variance Table

Model 1: psqi ~ vas.smerter
Model 2: psqi ~ gender + vas.smerter
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    361 5154.5
2    360 5097.4  1    57.097 4.0325 0.04538 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> anova(model5,model4)

Analysis of Variance Table

Model 1: psqi ~ vas.smerter
Model 2: psqi ~ depfac + vas.smerter
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    361 5154.5
2    360 5051.0  1    103.42 7.3712 0.006947 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(smodel2,smodel1)

Analysis of Variance Table

Model 1: sqrt(psqi) ~ gender + depfac + vas.smerter
Model 2: sqrt(psqi) ~ gender * depfac + vas.smerter
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    359 171.08
2    358 170.08  1   0.99489 2.0941 0.1487

> anova(smodel3,smodel2)

Analysis of Variance Table

Model 1: sqrt(psqi) ~ gender + vas.smerter
Model 2: sqrt(psqi) ~ gender + depfac + vas.smerter
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    360 174.55
2    359 171.08  1    3.4724 7.2867 0.007275 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(smodel4,smodel2)

Analysis of Variance Table

Model 1: sqrt(psqi) ~ depfac + vas.smerter
Model 2: sqrt(psqi) ~ gender + depfac + vas.smerter
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    360 172.75
2    359 171.08  1     1.672 3.5087 0.06186 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> anova(smodel5,smodel3)

Analysis of Variance Table

Model 1: sqrt(psqi) ~ vas.smerter
Model 2: sqrt(psqi) ~ gender + vas.smerter
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    361 176.44
2    360 174.55  1    1.8845 3.8866 0.04944 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(smodel5,smodel4)

Analysis of Variance Table

Model 1: sqrt(psqi) ~ vas.smerter
Model 2: sqrt(psqi) ~ depfac + vas.smerter
  Res.Df    RSS Df Sum of Sq     F   Pr(>F)
1    361 176.44
2    360 172.75  1    3.6848 7.679 0.005877 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> ### dele af summary() og confint() på udvalgte modeller:
> summary(model2)

               Estimate  Std. Error   t value     Pr(>|t|)
(Intercept)  5.41490382 0.357794958 15.134098 2.279144e-40
genderMand  -0.95396258 0.499023445 -1.911659 5.671585e-02
depfac1      1.69915904 0.643116723  2.642069 8.600671e-03
vas.smerter  0.06353701 0.007604744  8.354918 1.449345e-15

Residual standard error: 3.732 on 359 degrees of freedom

> confint(model2,conf.int=0.95)

                  2.5 %     97.5 %
(Intercept)  4.71126643 6.11854122
genderMand  -1.93533906 0.02741390
depfac1      0.43440959 2.96390849
vas.smerter  0.04858157 0.07849246
```

```
> summary(model3)

               Estimate  Std. Error   t value     Pr(>|t|)
(Intercept)  5.54783527 0.357170043 15.532756 5.276989e-42
genderMand  -1.00948504 0.502705064 -2.008106 4.537864e-02
vas.smerter  0.06518244 0.007641896  8.529615 4.129751e-16

Residual standard error: 3.763 on 360 degrees of freedom

> confint(model3,conf.int=0.95)

                  2.5 %     97.5 %
(Intercept)  4.84543342  6.2502371
genderMand  -1.99809249 -0.0208776
vas.smerter  0.05015407  0.0802108

> summary(model4)

              Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) 5.12153247 0.324401325 15.787644 4.865697e-43
depfac1     1.75093209 0.644910756  2.714999 6.946855e-03
vas.smerter 0.06635549 0.007487909  8.861685 3.683858e-17

Residual standard error: 3.746 on 360 degrees of freedom

> confint(model4,conf.int=0.95)

                 2.5 %     97.5 %
(Intercept) 4.48357278 5.75949215
depfac1     0.48266642 3.01919775
vas.smerter 0.05162996 0.08108103

> summary(model5)

              Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) 5.24113136 0.324220234 16.165343 1.319616e-44
vas.smerter 0.06822341 0.007521744  9.070158 7.768066e-18

Residual standard error: 3.779 on 361 degrees of freedom

> confint(model5,conf.int=0.95)

                 2.5 %     97.5 %
(Intercept) 4.60353376 5.87872895
vas.smerter 0.05343147 0.08301534
```

```
> summary(smodel2)

               Estimate  Std. Error   t value      Pr(>|t|)
(Intercept)  2.22900385 0.066182073 33.679874 3.645056e-113
genderMand  -0.17290110 0.092305397 -1.873142  6.186118e-02
depfac1      0.32111503 0.118958629  2.699384  7.275393e-03
vas.smerter  0.01206286 0.001406665  8.575499  2.992534e-16

Residual standard error: 0.6903 on 359 degrees of freedom

> confint(smodel2,conf.int=0.95)

                  2.5 %      97.5 %
(Intercept)  2.098850589 2.359157114
genderMand  -0.354428339 0.008626131
depfac1      0.087171713 0.555058349
vas.smerter  0.009296518 0.014829196

> summary(smodel3)

               Estimate  Std. Error   t value      Pr(>|t|)
(Intercept)  2.25412586 0.066094107 34.104793 8.377431e-115
genderMand  -0.18339400 0.093025278 -1.971443  4.943868e-02
vas.smerter  0.01237382 0.001414128  8.750137  8.350107e-17

Residual standard error: 0.6963 on 360 degrees of freedom

> confint(smodel3,conf.int=0.95)

                  2.5 %        97.5 %
(Intercept)  2.124146809  2.3841049076
genderMand  -0.366335227 -0.0004527705
vas.smerter  0.009592827  0.0151548075

> summary(smodel4)

              Estimate  Std. Error   t value      Pr(>|t|)
(Intercept) 2.17583171 0.059993122 36.268020 2.952866e-122
depfac1     0.33049865 0.119266496  2.771094  5.876846e-03
vas.smerter 0.01257369 0.001384776  9.079950  7.293463e-18

Residual standard error: 0.6927 on 360 degrees of freedom
```

```
> confint(smodel4,conf.int=0.95)

                 2.5 %     97.5 %
(Intercept) 2.057850712 2.29381271
depfac1     0.095952081 0.56504521
vas.smerter 0.009850427 0.01529696

> summary(smodel5)

              Estimate  Std. Error   t value      Pr(>|t|)
(Intercept) 2.19840670 0.059984740 36.649433 1.036862e-123
vas.smerter 0.01292627 0.001391615  9.288682  1.499246e-18

Residual standard error: 0.6991 on 361 degrees of freedom

> confint(smodel5,conf.int=0.95)

                2.5 %     97.5 %
(Intercept) 2.08044329 2.31637012
vas.smerter 0.01018958 0.01566296

> ### estimable() anvendt på udvalgte modeller:
> library(gmodels)
> est1<-c(0,1,30)
> est2<-c(1,1,30)
> est3<-c(0,1,sqrt(30))
> est4<-c(1,1,sqrt(30))
> est<-rbind(est1,est2,est3,est4)
> estimable(model3,est,conf.int=0.95)

       Estimate Std. Error   t value  DF Pr(>|t|)  Lower.CI  Upper.CI
est1  0.9459881  0.5924080  1.596852 360 0.111176 -0.219027 2.1110032
est2  6.4938233  0.4430489 14.657126 360 0.000000  5.622534 7.3651124
est3 -0.6524661  0.5126438 -1.272748 360 0.203929 -1.660619 0.3556865
est4  4.8953691  0.4618698 10.599024 360 0.000000  3.987067 5.8036710

> estimable(model4,est,conf.int=0.95)

     Estimate Std. Error   t value  DF     Pr(>|t|)  Lower.CI  Upper.CI
est1 3.741597  0.6631365  5.642273 360 3.402195e-08 2.4374890  5.045705
est2 8.863129  0.6133848 14.449541 360 0.000000e+00 7.6568618 10.069397
est3 2.114376  0.6424418  3.291156 360 1.096447e-03 0.8509658  3.377786
est4 7.235909  0.6638634 10.899696 360 0.000000e+00 5.9303712  8.541446
```

```
> estimable(smodel3,est,conf.int=0.95)

       Estimate Std. Error   t value  DF   Pr(>|t|)    Lower.CI   Upper.CI
est1  0.1878205 0.10962476  1.713304 360 0.08751758 -0.02776485 0.40340588
est2  2.4419464 0.08198594 29.784941 360 0.00000000  2.28071484 2.60317791
est3 -0.1156198 0.09486443 -1.218790 360 0.22372197 -0.30217787 0.07093825
est4  2.1385060 0.08546874 25.020916 360 0.00000000  1.97042533 2.30658677

> estimable(smodel4,est,conf.int=0.95)

      Estimate Std. Error   t value  DF     Pr(>|t|)  Lower.CI  Upper.CI
est1 0.7077094  0.1226371  5.770763 360 1.703310e-08 0.4665344 0.9488845
est2 2.8835412  0.1134363 25.419926 360 0.000000e+00 2.6604602 3.1066221
est3 0.3993676  0.1188099  3.361400 360 8.586734e-04 0.1657190 0.6330162
est4 2.5751993  0.1227715 20.975548 360 0.000000e+00 2.3337599 2.8166387
```