Kapitel 12

Lineære normalfordelingsmodeller i praksis

Når man begynder for alvor at anvende statistiske modeller dukker der en del begreber og problemer op, som den teoretiske fremstilling ikke tager højde for. For de lineære modellers vedkommende vil vi diskutere nogle af disse på baggrund af en analyse af Toyota Hiace datasættet ved hjælp af ISU.

ISUW (Windows versionen af ISU) er en lille statistikpakke til Windows maskiner, skrevet i Delphi på basis af Pascal procedurer udviklet (oprindeligt til DOS maskiner) fra omkring 1990. Programmet kan hentes fra forfatterens hjemmeside `www.mes.cbs.dk/~sttt`. Windows versionen fylder ikke meget (under 2 MB), og kan selvfølgelig heller ikke det samme som de store programsystemer til statistisk databehandling (SAS, GENSTAT, S+ osv). Men den er lettere at sætte sig ind i, og den kan håndtere alle modeller som er nævnt i disse forelæsningsnoter — og mange flere.

Man skal være opmærksom på, at selv om det følgende kan læses som en introduktion til ISU, så er det væsentlige indhold af mere generel karakter, og ganske uafhængigt af hvilke hjælpemidler man bruger til udregningerne. Den rækkefølge, hvori tingene foregår, strukturen af de operationer der foretages, fortolkningen af output etc. har ikke noget at gøre med hvilken statistikpakke man bruger. Selve kommandosyntaksen behøver man ikke at hæfte sig ved, hvis man ikke har tænkt sig at bruge ISU.

12.1. Indlæsning af data.

Tekstfilen `HIACE.TXT`, som indeholder det fulde datasæt (vi har tidligere set en del af det som eksempel 8.1, side 71) er gengivet nedenfor. Bemærk den forklarende tekst i slutningen af filen. Ved at anbringe den der undgår man at genere indlæsningen. Denne fil antages at være til stede på det bibliotek, hvorfra vi har startet en ISU session.

```
 76 82 14800 *  g
122 83 33000 d  g
125 86 29900 b  g
120 84 29800 b  g
 97 86 34500 d  g
 98 84 37000 d  g
120 86 40800 d  g
135 81  7000 d  g
 98 84 19800 d  g
 93 85 29900 b  g
```

```
102 84 22800 d  g
132 86 34500 d  g
 92 88 59800 d  g
 68 87 54800 d  g
 92 87 52800 d  g
 90 86 42800 d  g
150 87 46800 d  n
 74 89 69800 d  n
111 86 35800 *  n
135 83 18200 *  n
130 85 34000 *  n
109 88 43800 d  n
 86 88 43800 d  n
 86 88 52800 d  n
 95 84 29900 b  n
124 86 29900 d  n
150 87 37800 d  n
122 84 19300 d  n

28 Synede Toyota Hiacer.
1.  km kørt (i 1000 km)
2.  årgang
3.  pris
4.  d=diesel, b=benzin, *=uoplyst
5.  Blå avis:  g=okt 91, n=sep 92
```

Det første problem vi møder består således i at indlæse disse data. Det gør man i ISU på følgende måde:

Vi får brug for tre vektorer af længde 28 — kaldet “variates” — til at gemme oplysningerne om km. kørt, årgang og pris. De oprettes ved

```
VARiate KM AAR PRIS 28
```

Vi har her antydet, ved passende brug af store og små bogstaver, hvorledes kommandoen kan forkortes. Dens fulde navn er i dette tilfælde `VARIATE`, men man kan nøjes med at skrive `VAR` (faktisk kan man nøjes med at skrive `V`). Navnene `KM`, `AAR` og `PRIS` på de tre variates vi opretter skal derimod (naturligvis) skrives fuldt ud. Variabelnavne i ISU må højst være 8 tegn lange, og de skal starte med et bogstav eller et understregningssymbol. De øvrige tegn kan også være cifre `(1234567890)`. Man kan bruge små eller store bogstaver efter behag, ISU skelner ikke mellem disse når det gælder variabelnavne.

Vi får også brug for to faktorer af længde 28 på 2 niveauer til at gemme oplysningerne om type (diesel eller benzin) og annonceringstidspunkt. De oprettes af kommandoen

```
FACtor TYPE AVIS 28 2
```

Nu er de fem vektorer oprettet, men de indeholder foreløbig lutter nuller. For at indlæse deres værdier fra datafilen åbner vi den:

```
OPENinfile HIACE.TXT
```

Herefter kan vi indlæse data med kommandoen

```
READ KM AAR PRIS TYPE=*,d,b AVIS=,g,n
```

Bemærk at de fem vektorer naturligvis skal stå i samme rækkefølge som i datafilens linier. Faktornavnene `TYPE` og `AVIS` er udvidet med "navnelister" som fortæller hvordan faktorniveauerne 0, 1 og 2 er kodet på filen. Bemærk at niveau 0 af faktoren `TYPE` får betydningen "uoplyst" — det er en konvention som i almindelighed kan anbefales for faktorer med "manglende værdier". Da faktoren `AVIS` ikke har manglende værdier er en kode for niveau 0 ikke nødvendig, derfor kommaet umiddelbart efter lighedstegnet.

Til kontrol af at data er læst korrekt ind kunne man her bruge kommandoen `LIST`. Det vil vi ikke bruge plads på her.

Vi får brug for bilernes aldre på annonceringstidspunktet. Kommandoen

```
compute ALDER=90+AVIS-AAR
```

opretter en variate `ALDER` med (en passende approksimation til) disse værdier. Bemærk at når en faktor (her `AVIS`) på denne måde indgår i en formel er det med sine heltalsværdier, her 1 eller 2. Når kommandoen `COMPUTE` her er skrevet med små bogstaver er det simpelthen fordi den kan udelades — i et ISU–program kan man nøjes med at skrive `ALDER=90+AVIS-AAR`, og ved interaktiv kørsel kan man skrive `COMPUTE` kommandoer ved at indlede kommandolinien med et mellemrum.

Nu har vi de data vi vil regne på. Forudseende gemmer vi dem i et ISU datasæt med navnet `HIACE` (filens fulde navn bliver `HIACE.SUD`) ved hjælp af kommandoen

```
SAVEdata HIACE
```

Det betyder, at hvis vi senere får lyst til at regne på de samme tal i en anden ISU session kan vi springe alle de indtil nu forklarede indledende manøvrer over og hente dem direkte ind ved at skrive `GET HIACE`.

12.2. Analyse af den lineære model.

Vi begynder med at estimere en lineær model hvor responsen `PRIS` påvirkes additivt af faktorerne `TYPE` og `AVIS`, og hvor `ALDER` og `KM` indgår som regressionsvariable. Vekselvirkninger ser vi bort fra på grund af datamaterialets beskedne størrelse. Først må vi imidlertid fjerne de observationer der ikke er fuldstændigt oplyst med hensyn til alle de variable der skal indgå i modellen. I dette tilfælde er det kun faktoren `TYPE` der ikke er fuldt oplyst, så vi "skjuler" de fire observationer for hvilke `TYPE` er uoplyst med kommandoen

```
EXCLUDELevel TYPE 0
```

Den nævnte model estimeres nu ved kommandoen

```
FITLINearnormal PRIS=1+ALDER+KM+TYPE+AVIS
```

Denne kommando resulterer (som kørslens første) i output der forevises på skærmen. Man kan undersøge dette output ved på sædvanlig måde at bruge cursor op/ned pilene og PageUp/PageDown. Når man har kikket færdig trykker man på enten Return eller Escape. Return bruges hvis man vil have output føjet til kørslens outputfil, som man så eventuelt senere kan printe ud. Escape bruges hvis man ikke vil gemme det foreviste output.

Output fra `FITLINEARNORMAL` ser i dette tilfælde sådan ud:

```
ANALYSIS OF VARIANCE TABLE
Square sums and F-tests for removal of terms, last first.

  Effect  D.F.         S.S.          M.S.            F          P
-------------------------------------------------------------------
CONSTANT     1  33982900417   33982900417     168.7301   0.000000
   ALDER     1   3552710579    3552710579      72.3983   0.000000
      KM     1    115830939     115830939       2.5239   0.127073
    TYPE     1     21637121      21637121       0.4593   0.505702
    AVIS     1     40116870      40116870       0.8450   0.369483
RESIDUAL    19    901994074      47473372
-------------------------------------------------------------------
   TOTAL    24  38615190000

24 observations, 5 parameters estimated.
Estimated variance and standard deviation for PRIS.
          47473372                6890.09
R-square = 0.8053
```

Vi får her, som forklaret i kapitel 11, foretaget en række F–tests for modelreduktioner svarende til at modelformlens led fjernes ét for ét bagfra. Altså, for eksempel: Testet for reduktion af modellen `1+ALDER+KM+TYPE` til `1+ALDER+KM` fører til en P–værdi på 0.505702. Modellens tre sidste led kan åbenbart fjernes, så vi ender med en simpel regression af `PRIS` på `ALDER`. Bemærk, at dette kun kunne lade sig gøre fordi vi på forhånd har skønnet at `ALDER` er den vigtigste forklarende variabel, og derfor har skrevet den lige efter konstantleddet.

Størrelsen $R^2 = 0.8053$ ("`R-square`" fra sidste outputlinie) kaldes modellens *forklaringsgrad* eller *determinationskoefficienten*. Det er en størrelse man ofte angiver som mål for, hvor god en model er til at forklare observationernes variation. Den er defineret som

$$R^2 = \frac{\mathrm{SSD}_y - \mathrm{SSD}_{\mathrm{res}}}{\mathrm{SSD}_y},$$

og angiver således den andel af den totale kvadratafvigelsessum SSD_y af observationerne omkring deres middelværdi, der er forklaret af modellen (idet resten, SSD_{res}, er den del der ikke kan forklares af modellen). En forklaringsgrad tæt på 1 betyder, at modellens fittede værdier ligger tæt ved selve observationerne, medens en forklaringsgrad tæt på 0 betyder, at modellen ikke er meget bedre end den primitive "kapitel 7–model" der "forklarer" observationerne som identisk fordelte.

Da faktoren `TYPE` ikke skal bruges mere kan vi nu inkludere de fire observationer vi gemte af vejen før. Kommandoen

```
INCLUDEALL
```

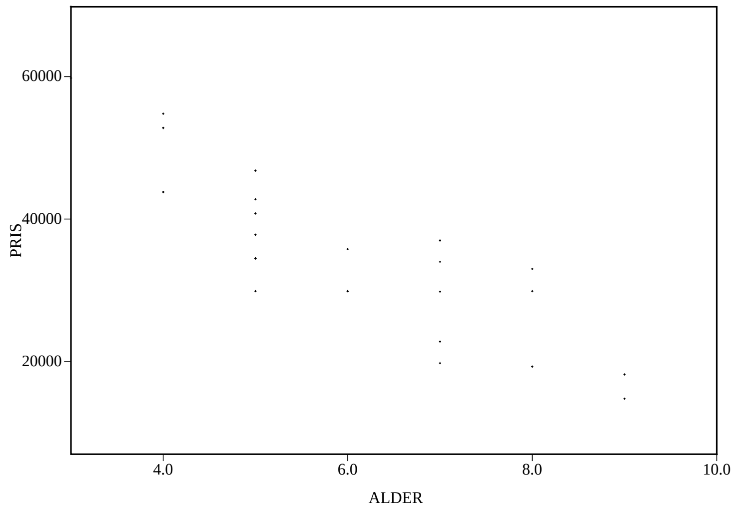
sletter alle tidligere indførte restriktioner.

Strengt taget burde vi nu undersøge om `AVIS` og `KM` stadig er insignifikante, når disse fire observationer er med. Men det tillader vi os at springe over.

For at vurdere om en simpel regression af `PRIS` på `ALDER` er en rimelig model laver vi den sædvanlige tegning:

```
PLOT ALDER PRIS
```

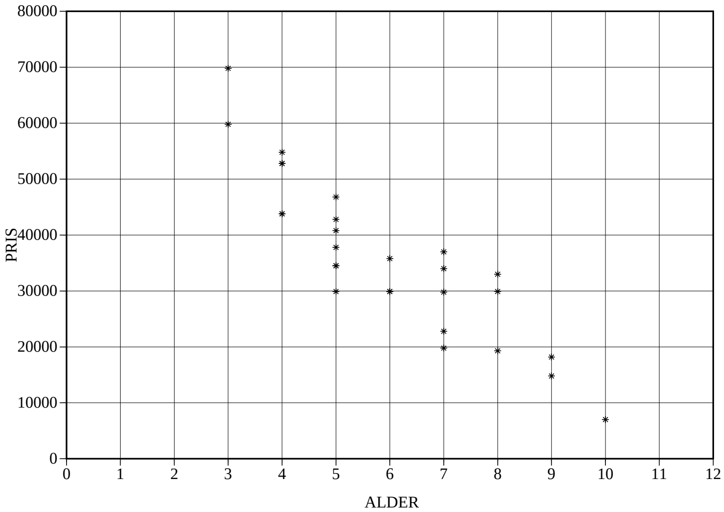
Kommandoen bevirker, at vi på skærmen får forevist en tegning der ser omtrent sådan ud:



Den er jo ikke specielt køn. Akseinddelingerne, især den lodrette, virker ret arbitrære, fordi rammen er valgt som det mindste rektangel der indeholder de 28 punkter. Der ligger faktisk et usynligt punkt i nederste højre hjørne. Men det ser ud som om vi ved at lade den vandrette akse gå fra 0 til (f.eks.) 12, og lade den lodrette gå fra 0 til 80000, vil få noget bedre. Vi prøver med

```
PLOT ALDER=0,-12,12 PRIS=0,-8,80000 =4 =*
```

Her bevirker `ALDER=0,-12,12` at den vandrette akse kommer til at gå fra 0 til 12 med markeringer svarende til 12 lige store delintervaller, og `PRIS=0,-8,80000` bevirker tilsvarende at den lodrette akse bliver delt i 8 lige store stykker fra 0 til 80000. Minustegnene foran 12 og 8 bevirker at der tegnes et gitter, med vandrette og lodrette linier i markeringspunkterne. Den tredje parameter `=4` (vi kunne også have skrevet `=red`) bevirker at punkterne på skærmen bliver røde (om ikke andet lidt sjovere end de sorte vi fik før), og den sidste parameter `=*` betyder at punkterne markeres med stjerner som er lidt større end de plusser man får hvis man ikke skriver noget. Resultatet ser sådan ud:



Hvis man tæller punkterne vil man se at der er 24. Det skyldes at 4 af dem er dobbeltpunkter. Mærkeligt nok er der fire par af biler som er annonceret som værende af samme alder og til samme pris. Hvilke det drejer sig om kan man se på tegningen side 72, hvor dobbeltpunkterne er vist ved at ændre lidt på alderen af den anden bil i hvert par. Denne tekniske detalje springer vi over her.

Da punkterne ser ud til at ligge pænt på linie fitter vi en simpel regressionsmodel ved

```
FITLINearnormal PRIS=1+ALDER
```

Denne kommando giver følgende output:

```
ANALYSIS OF VARIANCE TABLE
Square sums and F-tests for removal of terms, last first.

  Effect  D.F.         S.S.          M.S.             F          P
------------------------------------------------------------------
CONSTANT     1  36136957500   36136957500      178.4981   0.000000
   ALDER     1   4343417194    4343417194      100.5837   0.000000
```

```
RESIDUAL    26    1122735306      43182127
------------------------------------------------------------------
   TOTAL    28   41603110000

28 observations, 2 parameters estimated.
Estimated variance and standard deviation for PRIS.
          43182127               6571.31
R-square = 0.7946
```

Man bemærker at R^2–størrelsen er faldet en smule. Det gør den naturligvis altid når man reducerer modellen, fordi en hvilken som helst indskrænkning af en model giver anledning til et mindre middelværdiunderrum og dermed en vektor af fittede værdier som har større afstand til observationsvektoren. Det er karakteristisk at faldet er ubetydeligt for modelreduktioner der kan godkendes. Men lige i dette tilfælde kunne vi for øvrigt godt have risikeret at den steg, fordi vi jo har ændret lidt på datasættet.

Bortset fra variansestimatet får vi ikke noget at vide her, som vi ikke vidste i forvejen fra den tidligere `FITLIN` kommando. Men det er naturligvis nødvendigt at estimere den rigtige model for at få de rigtige parameterestimater skrevet ud. Det får vi nu, med kommandoen

```
LISTParameters
```

som generelt udskriver parameterestimater for den sidst fittede model. I dette tilfælde giver den følgende output:

```
                      Estimate        Std.dev.            T           P
CONSTANT                 75832            4168       18.192    0.000000
   ALDER                 -6731           671.2      -10.029    0.000000
```

Søjlen `Estimate` indeholder estimaterne for afskæring og hældning, anden søjle `Std.dev.` angiver disses estimerede standardafvigelser. De to sidste søjler indeholder T–teststørrelser og og tilhørende P–værdier for test af hypoteserne “parameter=0”. Disse testresultater udskrives altid, selvom de langtfra altid er relevante. I dette tilfælde får vi udført testet for hældning=0 (som vi har set før, i F–fordelingsvarianten) og testet for afskæring=0 (som er helt irrelevant i dette tilfælde).

12.3. Residualplottet.

Med henblik på blandt andet at lave en tegning, hvor regressionslinien også er indtegnet, skal vi have fat i de fittede værdier. Vi trækker med det samme residualerne ud (selv om vi næsten lige så godt kunne beregne dem bagefter ved `COMPUTE RES=PRIS-FIT`):

```
SAVEFitted FIT RES
```

Kommandoen `SAVEFITTED` kan have op til tre parametre, som skal være navne der ikke er i brug, eller navne på eksisterende variates af samme længde som de variates og/eller faktorer der indgik i sidste `FIT...` kommando. Effekten af kommandoen er, at de tre variates kommer til at indeholde henholdsvis fittede værdier, residualer og normerede residualer fra den sidst fittede model.

Et standard redskab til modelkontrol i forbindelse med lineære modeller er *residualplottet*, som er et plot af residualer mod fittede værdier. På dette plot kan man blandt andet se om der er tegn på at variansen vokser eller aftager med middelværdien, og undertiden kan man se tegn på andre afvigelser fra modellen der ytrer sig som krumning eller S–form. Tegningen er lidt overflødig i forbindelse med simpel regression, hvor man næsten lige så godt kan se direkte på det sædvanlige "x–y–plot". Men vi laver den alligevel, for illustrationens skyld.

For at få en pæn tegning undersøger vi først hvad variationsområdet for residualerne er, med kommandoen

```
SUMmary RES
```

Denne kommando giver følgende output:

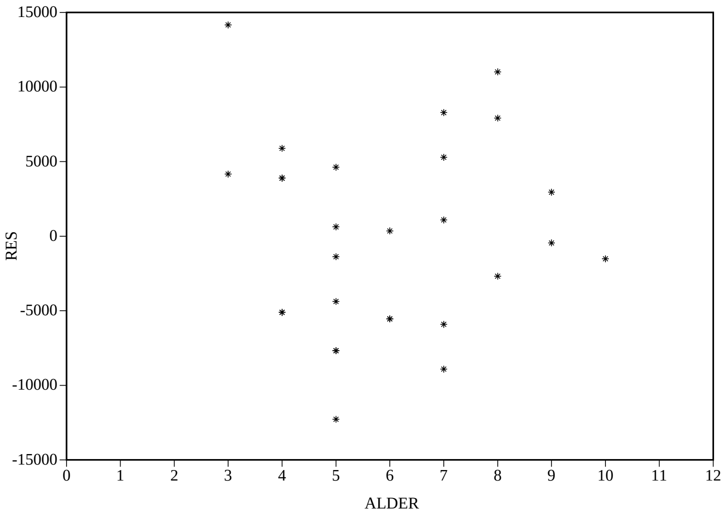
```
Variate RES, length 28

              present   excluded      total
non-missing        28          0         28
    missing         0          0          0
      total        28          0         28

Statistics for values present and non-missing.
Min  = -12275.55859375000
Mean =      -0.00003814697
Max  =  14161.69921875000
S.D. =   6448.47188572923
```

Vi får her blandt andet at vide at et passende symmetrisk interval som indeholder alle residualer er intervallet fra -15000 til 15000. Det område kan vi passende dele i 6 intervaller af længde 5000. I stedet for at bruge de fittede værdier som vandrette koordinater vælger vi her at bruge x–værdierne, altså værdierne af `ALDER`. Det er jo ligegyldigt, da `FIT` og `ALDER` er ens på nær affin transformation. Men denne simplifikation har naturligvis kun mening for en simpel regressionsmodel (altså en model med kun én regressionsvariabel). Vi inddeler den vandrette akse på samme måde som tidligere, men udelader denne gang minustegnene (dvs. vi vil ikke have tegnet et gitter):

```
PLOT ALDER=0,12,12 RES=-15000,6,15000 =0 =*
```



Der er ikke noget mistænkeligt ved dette residualplot. Man kan måske få en svag mistanke om at punktskyen krummer som en (meget lidt) glad parabel. Det vender vi tilbage til.

Det oprindelige plot af `PRIS` mod `ALDER` med regressionslinien indtegnet kan vi nu producere på følgende måde:

```
PLOT ALDER=0,12,12|ALDER PRIS=0,-8,80000|FIT =4|=15 =*|=L
```

Denne kommando skal læses som en "sammenfletning" af de to kommandoer

```
PLOT ALDER=0,12,12 PRIS=0,-8,80000 =4 =*
```

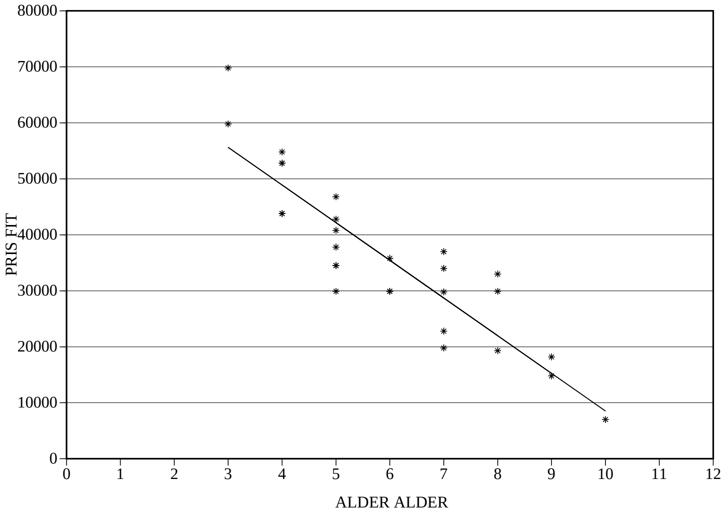
og

```
PLOT ALDER FIT =15 =L
```

De lodrette streger der bruges som skilletegn ("pseudoblanks" i ISU's jargon) kan under interaktiv kørsel og ved editering i ISUW's egen editor skrives ved hjælp af "1/2–tasten" (lige under Esc).

Den første af de to kommandoer, som her er flettet sammen (hvilket betyder at man får de to plots tegnet oven i hinanden) giver bare, på nær uvæsentlige detaljer, det sædvanlige x–y punktplot. Den anden plotter fittede værdier mod alder, hvilket umiddelbart fører til at visse punkter på regressionslinien markeres. Men den sidste parameter `=L` (som generelt bestemmer plottesymbolet) bevirker at punkterne forbindes med linier. Tredje parameter (farveparameteren) `=15` (vi kunne også have skrevet `=white`) bevirker at selve punkterne ikke tegnes. Bemærk at den sidste af de to sammenflettede kommandoer ikke indeholder specifikationer af akserne. Hvis de var der ville de blive ignoreret. Rammen skal være den samme for de to overlejrede tegninger, og angives derfor kun for den første.

Resultatet ser sådan ud:



x–y–plot med indtegnet regressionslinie

12.4. Modelkontrol.

Når man i en simpel regressionsmodel har flere y'er for hvert x (her: priser på flere biler af samme alder) ligger det lige for at kontrollere regressionsmodellen ved test imod den ensidede variansanalysemodel, der tilordner hver af de forekommende x–værdier sin egen frit varierende middelværdi. Det vil i dette tilfælde sige en model, hvor priserne antages uafhængige, normalfordelte med samme varians og en middelværdi der afhænger på helt vilkårlig måde af alderen. Det svarer blot til at vi opfatter alderen som en kvalitativ variabel, altså en gruppering, hvor selve aldrene ikke tillæges anden betydning end navne, der er hæftet på grupperne.

Da der forekommer 8 aldre 3, 4, ... , 10 får vi brug for en faktor på 8 niveauer:

```
FACtor ALDGRP 28 8
```

Vi giver den de rigtige værdier fra 1 til 8 ved

```
compute ALDGRP=ALDER-2
```

Niveauet `ALDGRP=1` svarer således til `ALDER=3`, osv. Den ensidede variansanalysemodel fittes nu ved

```
FITLINearnormal PRIS=ALDGRP
```

som giver følgende output:

```
ANALYSIS OF VARIANCE TABLE
Square sums and F-tests for removal of terms, last first.

  Effect  D.F.          S.S.          M.S.            F          P
--------------------------------------------------------------------
  ALDGRP     8   40895598190    5111949774      144.5050   0.000000
RESIDUAL    20     707511810      35375590
--------------------------------------------------------------------
   TOTAL    28   41603110000
28 observations, 8 parameters estimated.

Estimated variance and standard deviation for PRIS.
          35375590                5947.74
R-square = 0.8706
```

Her kunne vi også have valgt at skrive `FITLIN PRIS=ALDER+ALDGRP`. Da en model med alderen både som faktor og som regressionsvariabel jo ikke er andet end den ensidede variansanalysemodel bestemt ved faktoren (bare overparametriseret på en komplet uoverskuelig måde), ville vi faktisk få det ønskede test ud i sidste linie af variansanalyseskemaet. Men her vælger vi i stedet at illustrere en anden ISU–facilitet. Kommandoen `TESTMODELCHANGE` foretager kvotienttest for den sidst fittede model imod den foregående. Og da den foregående netop var regressionsmodellen giver kommandoen

```
TESTmodelchange
```

det ønskede test. Output ser sådan ud:

```
6 parameters added
variance-ratio =            1.956262
P[ F(6,20) > variance-ratio ] =  0.120689
```

OBS: kommandoen `TESTMODELCHANGE` er farlig derved at den kræver at de to sidst fittede modeller skal stå i en ganske bestemt relation til hinanden. Den ene skal være en delmodel af den anden, fittet på de samme data osv. Det er ISU–brugerens eget ansvar at sørge for at dette er i orden, ISU har (næsten) ikke mulighed for at kontrollere det.

Men konklusionen af testet er altså, at den ensidede variansanalysemodel ikke er en signifikant forbedring af regressionsmodellen. Og det betyder naturligvis at regressionsmodellen må foretrækkes, da den på en langt mere effektiv måde tager hensyn til problemets struktur.

Man kan også mere specifikt undersøge om det har noget på sig, at punktskyen krummer lidt for meget opad til at kunne beskrives med en linie. Almindelig sund fornuft taler jo for, at hvis datamaterialet havde omfattet nye eller næsten nye biler, så ville vi ikke kunne bruge en lineær model. Måske ville vi her få noget, der for nyere biler bedre kunne beskrives ved en fast procentvis afskrivning, altså en faldende eksponentiel kurve, og det kunne godt tænkes at slå svagt igennem også efter tre år. Hvis der er en sådan krumning må man formode, at hvis vi fitter en model der beskriver prisen som et andengradspolynomium i alderen, så vil et test for linearitet i denne model (altså for hypotesen om at andengradsleddet er 0) afsløre det. Det kunne vi gøre ved at tilføje en variabel til modellen, hvis værdier var de kvadrerede aldre (`ALDER2=SQR(ALDER)`, derefter `FITLIN PRIS=1+ALDER+ALDER2`). Men modelformelsyntaksen tillader også (som antydet side 110 og 111) at vi simpelthen skriver

```
FITLINearnormal PRIS=1+ALDER+ALDER*ALDER
```

Denne kommando producerer følgende variansanalyseskema:

```
     Effect  D.F.          S.S.          M.S.             F          P
---------------------------------------------------------------------
   CONSTANT     1   36136957500   36136957500      178.4981   0.000000
      ALDER     1    4343417194    4343417194      100.5837   0.000000
ALDER*ALDER     1     126255909     126255909        3.1675   0.087275
   RESIDUAL    25     996479396      39859176
---------------------------------------------------------------------
      TOTAL    28   41603110000
28 observations, 3 parameters estimated.
```

Vi får altså hypotesen om linearitet godkendt med en P–værdi på 8.7%.

12.5. Multikollinearitet.

Lad os til sidst prøve noget lettere vanvittigt: Hvor godt kan vi forudsige priserne *alene* på basis af variablen `KM`, som vi tidligere har testet væk? Vi prøver med

```
FITLIN pris=1+km
```

som giver følgende output:

```
ANALYSIS OF VARIANCE TABLE
Square sums and F-tests for removal of terms, last first.

  Effect  D.F.          S.S.          M.S.             F          P
---------------------------------------------------------------------
CONSTANT     1   36136957500   36136957500      178.4981   0.000000
      KM     1     858158454     858158454        4.8420   0.036857
RESIDUAL    26    4607994046     177230540
---------------------------------------------------------------------
   TOTAL    28   41603110000

28 observations, 2 parameters estimated.
Estimated variance and standard deviation for PRIS.
         177230540               13312.8
R-square = 0.1570
```

Vi ser her, først og fremmest, at modellen har en meget lavere forklaringsgrad ($R^2 = 0.157$) end modellen med alder som forklarende variabel. Hvis man laver det tilsvarende "x–y–plot" vil man da også se, at der stort set ikke er nogen sammenhæng mellem `KM` og `PRIS`. Men det interessante er, at hældningen er svagt signifikant (P=0.037). Den viser sig at være negativ, svarende til at prisen falder med ca. 250 kr. pr. 1000 km kørt. Dette, at en forklarende variabel (`KM`) pludselig går hen og bliver signifikant fordi vi fjerner en anden forklarende variabel (`ALDER`), kan forklares ved et fænomen som kaldes *multikollinearitet*, som opstår når to eller flere baggrundsvariable er korrelerede. I dette tilfælde er der en svag korrelation mellem `ALDER` og `KM`, som naturligvis skyldes, at jo ældre en bil er, jo længere har den i almindelighed kørt. I betragtning af denne logiske sammenhæng er korrelationen i øvrigt forbavsende svag (korrelationskoefficienten er 0.33, og den er ikke engang signifikant forskellig fra 0). Men det betyder, at `KM` har en svag tilbøjlighed til at følges med `ALDER`, og når `ALDER` ikke er til stede i modellen overtager `KM` en del af `ALDER`'s forklaringsevne.

I mere ondskabsfulde tilfælde af multikollinaritet kan man opleve følgende: I en model med to forklarende variable kan begge testes væk hver for sig, men så snart den ene er fjernet bliver den anden ekstremt signifikant. Det kunne let være sket i Toyota Hiace eksemplet under andre teknologiske omstændigheder. Når alderen er så vigtig og kilometertallet så uvigtigt er det naturligvis fordi en bils levetid her i landet i det væsentlige er karosseriets levetid, som ikke afhænger ret meget af hvor meget bilen bliver kørt. Hvis motorer var lidt mere sarte eller karosserier lidt mere robuste, kunne `KM` have haft lige så stor indflydelse på prisen som `ALDER`. Og hvis det yderligere var sådan (som det altså ikke er), at `KM` og `ALDER` var stærkt korrelerede, så kunne man være havnet i en situation som den beskrevne, hvor hver af de to variable kunne bruges til en ret effektiv forudsigelse af prisen, men hvor man ikke ville få nogen væsentlig forbedring ved at inddrage den anden også.

`Quit`