S-PLUS 4.5 voor Windows

Mathisca de Gunst[1]

Naam: S-PLUS for Windows.
Versienummer: 4.5 release 2.
Uitgever: MathSoft. Inc., Seattle, Washington.
Importeur: CANdiensten, Amsterdam; e-mail: info@candiensten.nl.
Prijs: *f* 6.000,00 voor een enkelvoudige S-PLUS licentie onder Windows.

Systeemvereisten:

- Minimale platform configuratie: 486 IBM compatibele PC (pentium aanbevolen), snelheid 90 MHz, 32MB RAM geheugen (64 aanbevolen).
- Vereist geheugen harde schijf: 80MB.
- Microsoft Windows 95 of Windows NT.
- CD-ROM.
- Windows compatibele printers worden ondersteund.
- Er zijn netwerklicenties beschikbaar voor TCP/IP of Novell Netware netwerken.

Inleiding

S-PLUS is een software-pakket voor statistische data analyse en de daarbij behorende grafische weergave. Het is gebaseerd op de programmeertaal S en heeft een interactieve programmeeromgeving. Het pakket is reeds een jaar of tien op de markt. Aanvankelijk werd het vooral in de academische wereld gebruikt, maar de afgelopen jaren heeft het een grotere bekendheid en wijder verspreid gebruik gekregen. De aanleiding voor deze bespreking is het uitkomen van S-PLUS 4.5 voor Windows.

In de versie voor Windows zijn naast enkele nieuwe statistische technieken en grafische mogelijkheden, vooral de grafische (Microsoft Office compatibele) user interface en de uitgebreide mogelijkheden voor data in- en uitvoer interessant.

Het pakket komt met een uitgebreide documentatie, drie dikke manuals—User's Guide, Programmer's Guide, Guide to Statistics—die ook on-line beschikbaar zijn, en een on-line Help menu. Inmiddels zijn er ook verschillende boeken verschenen over het gebruik van S-PLUS, onder andere over de taal S-PLUS ([1], [7]), over statistische data analyse met S-PLUS in het algemeen ([3], [8] (besproken in KM 58)) en voor specifieke toepassingen ([2], [4], [6]). Bovendien is er allerlei informatie beschikbaar via Internet.

Statistische functies

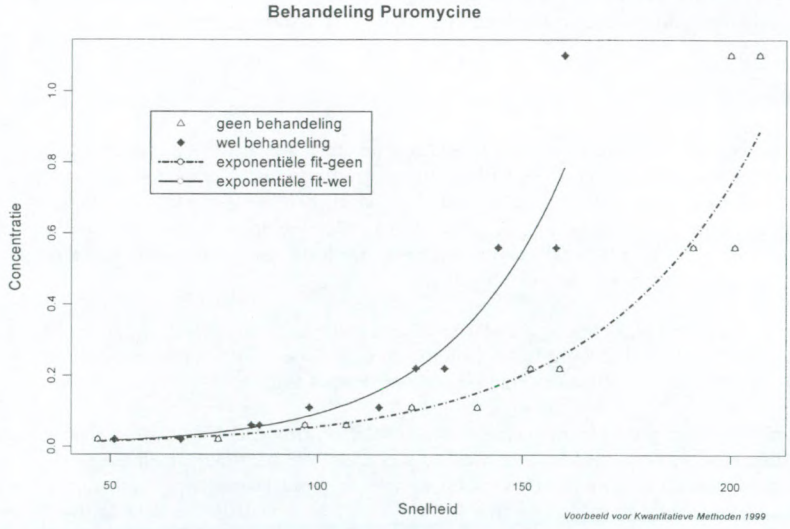
In S-PLUS kunnen statistische analyses op elk gewenst niveau worden uitgevoerd. Er zijn functies voor het samenvatten van data, zowel grafisch (histogram, boxplot, stem-and-leaf plot, qq-plots, etc.) als numeriek (gemiddelde, kwantielen, variantie e.d.). De meeste

[1] Vrije Universiteit, Divisie Wiskunde en Informatica, Faculteit der Exacte Wetenschappen.

standaard toetsen kunnen worden uitgevoerd en verschillende klassen van modellen kunnen worden gefit. Het pakket bevat onder meer routines voor het fitten van lineaire, niet-lineaire en gegeneraliseerde regressiemodellen (waaronder mixed effect modellen), voor variantie-analyse, voor multivariate technieken als factoranalyse en principale componentenanalyse, voor tijdreeksanalyse, survivalanalyse en quality control charts. Robuuste methoden, smoothing en resampling technieken staan ook ter beschikking. Voor het uitvoeren van analyses betreffende design en analyse van industriële experimenten, GARCH modellen voor financiële tijdreeksen, wavelets, optimalisatie of ruimtelijke statistiek, dient men aparte modules aan te schaffen (de laatste module is besproken in KM 58).

Grafische mogelijkheden

S-PLUS biedt de gebruiker een uitgebreid scala aan grafische mogelijkheden. Naast grafische samenvattingen van data, kunnen de resultaten van bovengenoemde statistische routines op de gebruikelijke wijze in beeld worden gebracht. Er kunnen, om slechts een paar voorbeelden te noemen, scatter plots, loess plots, trellis plots, time plots en contour plots worden gemaakt.



Figuur 1 Voorbeeld Graph sheet (embedded in Word)

De plaatjes kunnen naar eigen inzicht bewerkt worden; punten, functies kunnen worden toegevoegd, de plaatjes kunnen van tekst en kleur worden voorzien, etc. In de Windows versie kunnen ze worden bewaard in S-PLUS of worden uitgevoerd naar andere formaten voor grafische afbeeldingen als WMF, BMP, GIF, TIFF en Postscript. Een S-PLUS grafiek kan ook via OLE in o.a. Microsoft Word of Powerpoint worden gebruikt.

Data

Data kunnen in S-PLUS in verschillende vormen worden opgeslagen en verwerkt. S-PLUS gebruikt, bijvoorbeeld, objecten van de klassen vector, matrix en dataframe (een speciaal object voor het analyseren van statistische modellen). Daarnaast kan een object, of een element hiervan, nog van een bepaald type zijn, zoals numeric, character, logical en list (vector met als elementen willekeurige S-objecten). Voor elke klasse en elk type van objecten zijn specifieke operaties gedefinieerd. Data kunnen simpelweg worden ingetypt, maar ook worden ingevoerd uit en uitgevoerd naar andere applicaties. Naast ASCII data bestanden kunnen Excel, Lotus, dBase, SPSS, SAS, Microsoft Access, Matlab en vele andere data bestanden worden ingelezen en S-PLUS data kunnen worden uitgevoerd naar dezelfde formaten.

De S taal

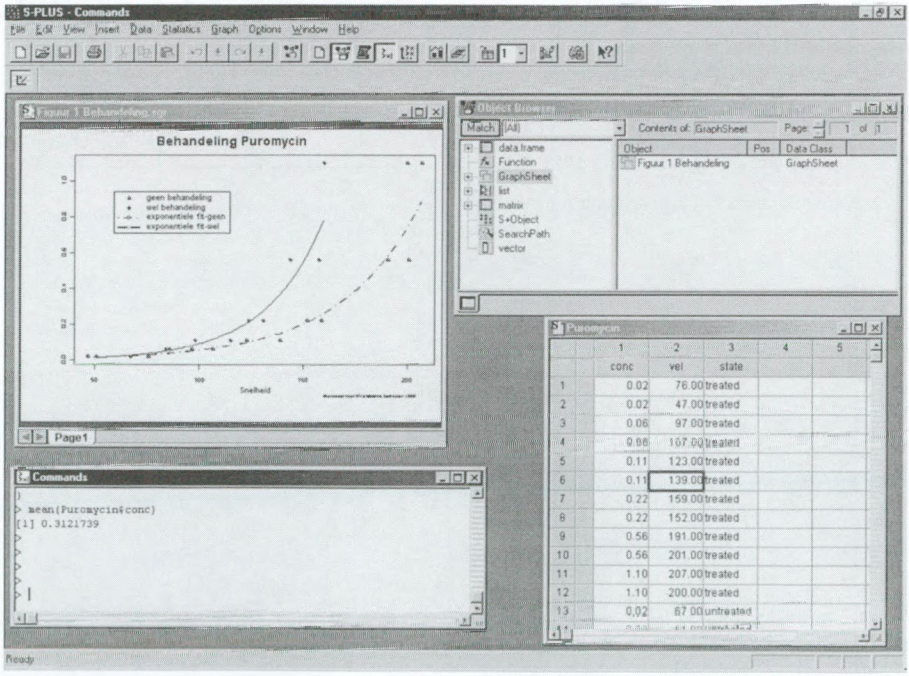
De S taal is een geïnterpreteerde, object-georiënteerde programmeertaal, ontwikkeld bij AT&T (nu Lucent Technologies). Men kan met S-PLUS een data analyse uitvoeren en grafieken produceren, zonder zelf enig programma te hoeven schrijven. Het schrijven van programma's in S is echter nogal eenvoudig, zodat bijna iedereen die S-PLUS regelmatig gebruikt, dit in de praktijk wel in meer of mindere mate zal doen. Bovendien is van elke standaard S-PLUS functie de definitie te zien door simpelweg de naam van de betreffende functie te typen. Deze beschikbaarheid van de S code stelt elke gebruiker in staat functies naar eigen inzicht aan te passen of te creëren.

```
> mean
function(x, trim = 0, na.rm = F)
{
        if(na.rm) {
                wnas <- which.na(x)
                if(length(wnas))
                        x <- x[ - wnas]
        }
        if(mode(x) == "complex") {
                if(trim > 0)
                        stop("trimming not allowed for complex data")
                return(sum(x)/length(x))
        }
        x <- as.double(x)
        if(trim > 0) {
                if(trim >= 0.5)
                        return(median(x, na.rm = F))
                if(!na.rm && length(which.na(x)))
                        return(NA)
                n <- length(x)
                i1 <- floor(trim * n) + 1
                i2 <- n - i1 + 1
                x <- sort(x, unique(c(i1, i2)))[i1:i2]
        }
        sum(x)/length(x)
>
```

Figuur 2 Code van de S-PLUS functie 'mean'

De gebruiker interface

De nieuwe, grafische user interface voor Windows maakt het uitvoeren van boven beschreven zaken redelijk eenvoudig. Zoals in andere Windows applicaties, wordt gewerkt met menu's, werkbalken en vensters. Het gebruik hiervan is niet anders dan bij andere Windows applicaties. Er zijn zes typen vensters: de Object browser, Data vensters, Graph sheets, het Commands venster, Script vensters en Report vensters. Sommige venster typen hebben hun eigen werkbalken en menu's. Hoewel men de omgeving naar eigen inzicht kan wijzigen, is de standaard dat bij het opstarten van S-PLUS alleen de Object browser en het Commands venster verschijnen. In het hoofdmenu dat dan actief is, staan onder meer de gebruikelijke File, Edit, View, Insert en Format menu's, alsook een menu voor het creëren en manipuleren van data, een menu waarmee de statistische functies kunnen worden aangeroepen, en een menu waarmee de grafieken worden gemaakt. Het hoofdmenu heeft ook een aantal knoppen in de werkbalk, bijvoorbeeld twee knoppen die pallets voor 2-, respectievelijk, 3-dimensionale grafieken tevoorschijn laten komen. Overigens is het eenvoudig mogelijk om de gehele user interface aan de eigen wensen aan te passen. Via selectie in de Object browser en achtereenvolgende keuzen in verschillende menu's kunnen menu's, werkbalken en dialoogvensters gecreëerd of aangepast worden.



Figuur 3 Voorbeeld S-PLUS scherm

De Object browser lijkt enigszins op de verkenner van Windows. Men kan hiermee S-PLUS objecten zoeken, selecteren, bekijken en wijzigen. Aangezien S-PLUS object-georiënteerd is, behoort elke systeemcomponent tot een klasse van objecten. Bij objecten dient men dan niet alleen te denken aan data-objecten, als vectoren, matrices en data frames, maar ook aan functies, grafische objecten, menu's, enz. In de Object browser staan de objecten geordend naar klasse. Net als bij de verschillende typen data kunnen op de objecten uit een zelfde klasse dezelfde soort handelingen worden verricht.

In een Data venster ziet men een data-object geordend in rijen en kolommen. Men kan in het venster (gedeelten van) het object wijzigen, gedeelten selecteren om vervolgens functies op de geselecteerde gedeelten toe te passen, rijen en kolommen toevoegen, uit een andere applicatie gegevens kopiëren en dergelijke. Voor het bewerken van een ander data-object wordt een nieuw Data venster geopend. Er kunnen dus meerdere Data vensters tegelijkertijd geopend zijn.

Een in S-PLUS gemaakte grafiek wordt in een Graph sheet gezet. Daarin kan het plaatje bewerkt worden. De Graph sheet kan desgewenst als object bewaard worden. Voor elke nieuwe grafiek wordt in principe een nieuwe Graph sheet geopend, maar men kan ook meerdere grafieken naast of in elkaar in een Graph sheet plaatsen.

Wie liever S-PLUS gebruikt door middel van het typen van commando's, zoals standaard is in de niet-Windows versies van S-PLUS, kan dit doen in het Commands venster. Achter de prompt in het venster kunnen interactief opdrachten in de S-PLUS taal getypt worden. Dit kan een berekening of functie opdracht zijn, waarna meteen de uitkomst of het resultaat op het scherm verschijnt; het kan een toekenning zijn, waarna een object in de werk directory datgene wat toegekend is bevat, of een grafische opdracht die vervolgens in een Graph sheet wordt uitgevoerd. Meer algemeen heeft men in dit venster toegang tot de S programmeertaal. Men kan hier dan ook bestaande functies wijzigen of nieuwe functies schrijven.

Ook via een Script venster heeft men toegang tot de S programmeertaal. In een Script venster kunnen scripts van S-PLUS commando's worden gecreëerd, bewerkt, geopend, opgeslagen en afgedrukt. Het venster bestaat uit twee delen: bovenin worden de opdrachten ingevoerd, onderin verschijnen de uitkomsten. Terwijl in het Commands venster een ingevoerde opdracht meteen wordt uitgevoerd, kunnen in een Script venster eerst verschillende opdrachten worden ingevoerd, waarna deze achter elkaar worden uitgevoerd. Een dergelijk venster is handig voor het schrijven van functies, het maken van menu's, knoppen en dialogen, en voor het automatiseren van herhaaldelijk terugkerende taken. Een gemaakt script kan men opslaan als een S-PLUS scriptbestand.

De uitvoer van een operatie die uitgevoerd is met behulp van een dialoog, komt in een Report venster te staan. De inhoud van een Report venster kan worden bewerkt en opgeslagen als een tekst of rtf bestand.

Omvang en snelheid

Voor de werkbare omvang van bestanden geldt voor de versie 4.5 voor Windows de vuistregel waarbij het RAM geheugen van de PC gedeeld door 4 de maximale omvang van het bestand in Mb is. Dus op een 64 Mb machine mag de data set 16 Mb groot zijn, ofwel

100.000 cases met 20 variablen (in doubles). Bij grotere bestanden gaat S-PLUS swappen en wordt dus erg langzaam. Om een indicatie te geven, is op een 256 Mb PentiumII-450 machine een 19 Mb dBase bestand (ruim 200.000 records) ingelezen. Dit ging zonder problemen in 1.5 minuut, waarna binnen 7 seconden 7 summary statistics en in 0.5 minuut een scatter plot van twee variabelen konden worden gemaakt. Er wordt wel geklaagd dat S-PLUS langzaam is bij complexe berekeningen via zelfgeschreven functies. Bij grote programma's is het inderdaad vaak sneller om in C of Fortran te programmeren en alleen de statistische en grafische resultaten uit S-PLUS te gebruiken. Soms hangt de traagheid echter samen met de wijze van programmeren. Een programma met veel loops is in S-PLUS doorgaans traag. Veelal kunnen loops in S-PLUS vermeden worden, bijvoorbeeld doordat een berekening op elementen van een matrix met een S-PLUS functie in een keer kan worden uitgevoerd en niet voor elk element apart hoeft te worden gedaan.

Tot slot

S-PLUS is in het algemeen redelijk gebruikersvriendelijk; de basisbeginselen zijn simpel. Op de universiteit hebben we er al jaren goede ervaringen mee in het statistiekonderwijs: de studenten kunnen er snel mee uit de voeten. Er kan heel veel met het pakket, zoals uit bovenstaande blijkt, maar men moet wel het nodige van statistiek weten om het goed te kunnen gebruiken. Je zou misschien kunnen zeggen, dat S-PLUS zich vooral op de statisticus richt, terwijl een pakket als SPSS van oorsprong meer bedoeld is voor statistische toepassingen binnen andere disciplines. Kenmerkend hiervoor is het bovengenoemde feit dat de in S-PLUS ingebouwde functies geheel open zijn. Door deze toegankelijkheid en de flexibiliteit van de programmeeromgeving ontwikkelt S-PLUS zich ook snel: er zijn regelmatig gebruikers die nieuwe statistische technieken ontwikkelen en de S-PLUS code daarvoor beschikbaar stellen. In nieuwe versies van het pakket zijn dan ook vaak recent ontwikkelde technieken al opgenomen.

De bijgeleverde documentatie is zeer uitgebreid, aardig volledig en in het algemeen goed leesbaar. Maar ook hiervoor geldt, dat je er geen statistiek mee kunt leren. De handleidingen zijn zeker niet overbodig, want hoewel gebruik van het pakket naar mijn idee gemakkelijk te leren is, kun je niet zonder meer het pakket opstarten en beginnen—ook niet met de versie voor Windows. Voor een snelle introductie in het pakket voor Windows kan men echter uitstekend terecht bij de "stoomcursus" [5], die speciaal voor dit doel is geschreven.

Voor mensen die regelmatig met grote data sets werken, al dan niet geïmporteerd uit andere applicaties, lijkt me de Windows versie nuttig. Voor hen die reeds vertrouwd zijn met een niet-Windows versie van S-PLUS, zijn mijns inziens aan het pakket voor Windows zowel voor- als nadelen verbonden. Sommige ingewikkelde operaties waarvoor bij gebruik van een niet-Windows versie de nodige opdrachten ingetypt dienen te worden, kunnen in de versie voor Windows met één druk op de knop gerealiseerd worden. Dit geldt met name voor het fitten van modellen en het produceren van grafieken met alles erop en eraan. Dat is heel prettig en kan veel tijd schelen. Een nadeel is dat er zóveel mogelijkheden zijn, dat het niet altijd duidelijk is waar te zoeken naar wat je nodig hebt. Weliswaar is zeer uitgebreide documentatie beschikbaar, maar met name in de User's guide staat veel over wat er allemaal kan, maar niet altijd over hoe het moet. Vaak is het dan gemakkelijker opdrachten in het Commands venster te typen, dan de menu's te gebruiken.

Tenslotte zijn er in deze versie nog wel enkele dingen die niet helemaal goed gaan. Bijvoorbeeld, bij het importeren van gegevens uit een MS Access database wordt automatisch alleen de eerst voorkomende tabel ingelezen, zonder mogelijkheid om een andere tabel te selecteren. Ook het lezen van MS Excel bestanden verloopt niet altijd vlekkeloos. Verder is lastig dat, na het laden van een module de bijbehorende functies wel beschikbaar zijn, maar uitsluitend in het Commands venster en niet via een menu. Je moet dus weten welke het zijn om ze te kunnen gebruiken. Maar dit zijn details vergeleken bij de veelheid aan statistische en grafische mogelijkheden die S-PLUS biedt.

Referenties

[1] Becker, R.A., Chambers, J.M., Wilks, A.R. (1988). The new S language. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA.
[2] Bruce, A., Gao, H.-Y. (1996). Applied wavelet analysis with S-PLUS. Springer-Verlag, New York.
[3] Chambers, J.M., Hastie, T.J. (eds.) (1992). Statistical models in S. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA.
[4] Härdle, W. (1991). Smoothing techniques with implementation in S. Springer-Verlag, New York.
[5] Lam, L.H. (1998). Stoomcursus S-PLUS voor Windows; grafische gebruikers interface. CANuitgeverij, Amsterdam.
[6] Marazzi, A. (1992). Algorithms, routines and S functions for robust statistics. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA.
[7] Spector, P. (1994). An introduction to S and S-PLUS. Duxbury Press, Belmont, CA.
[8] Venables, W.N., Ripley, B.D. 2nd ed. (1997). Modern applied statistics with S-PLUS. Springer-Verlag, New York.