İSTATİSTİKSEL YAZILIM GELİŞTİRME ORTAMI: R

A. Fırat ÖZDEMİR, Engin YILDIZTEPE, Mustafa BİNAR
Dokuz Eylül Üniversitesi, Fen-Edebiyat Fakültesi, İstatistik Bölümü

AKADEMİK BİLİŞİM 2010
10-12 Şubat 2010 MUĞLA ÜNİVERSİTESİ

İÇERİK

1. Giriş
2. R istatistiksel programlama dili
    1. R'de Nesne Kavramı Ve Nesneler
    2. R'de Nesne Olarak Fonksiyonlar
    3. R Kaynak Kod Dosyası
    4. Çöp Toplayıcı (Garbage Collector)
    5. Yardımın Kullanılması
3. Sonuç

1-GİRİŞ

- R dili ilk olarak Yeni Zelanda'daki Aucland Üniversitesi İstatistik Bölümü'nden Ross Ihaka ve Robert Gentleman tarafından yazılmıştır.
- Daha sonra dünyanın çeşitli yerlerindeki araştırmacılar R' yi geliştirmek için bir araya gelmiş ve 1997'de bu gruba "**R core team**" adı verilmiştir.
- R diline, ilk geliştiricileri olan **R**oss Ihaka ve **R**obert Gentleman tarafından S diline atıfta bulunularak "R" ismi verilmiştir.

1-GİRİŞ

- R dilinin tasarımı önemli ölçüde Becker, Chamber, ve Wilks'in geliştirdiği S dili ile Sussman'ın geliştirdiği Scheme dillerinden etkilenmiştir.

- Görünüm özellikleri açısından S diline benzeyen R, uygulama ve anlamsal yönden Scheme diline yakındır.

- S-PLUS yazılımının akademik ve öğretim amaçlı kullanılmasında lisans ücretlerinin pahalı bulunması nedeniyle Yeni Zelanda'lı iki istatistikçi Ross Ihaka ve Robert Gentleman, "R" adını verdikleri programlama dilini geliştirmeye karar vermişlerdir.

- R dilinin ilk sürümü "R core team" tarafından 29 Şubat 2000 tarihinde yayınlanmıştır.

1-GİRİŞ

- R istatistiksel yazılım geliştirme ortamı veri manipülasyonu, hesaplama ve grafik gösterim için tasarlanmıştır.

- R dilinin söz dizimi kuralları (syntax) C diline benzerlik gösterir. Fonksiyonel bir programlama dili olan R istatistikçiler ve matematikçiler için kod yazmayı kolaylaştıran fonksiyonlara sahiptir.

- Bu çalışmada açık kaynak kodlu ve ücretsiz bir programlama dili olan R, tarihsel gelişimi, nesne kavramı, fonksiyonları, yardım seçenekleri ve diğer özellikleri ile incelenmiştir.

2-R İSTATİSTİKSEL PROGRAMLAMA DİLİ

- R-Project'in web sitesinde yapılan tanıma göre R, istatistiksel hesaplamalar ve grafikler için bir **dil ve ortamdır.**
- R, yaygın olarak kullanılan SPSS, SAS gibi istatistik paket programlardan farklıdır. R **bir istatistik paket program değil istatistiksel yazılım geliştirme ortamıdır.**

2-R İSTATİSTİKSEL PROGRAMLAMA DİLİ ÖZELLİKLERİ

- Etkin veri işleme ve saklama özelliğine sahiptir.
- Dizi ve özellikle matris hesaplamalarında kullanılabilecek **özel operatörler** mevcuttur.
- Veri analizi için kullanılabilecek **uyumlu ve bir arada kullanılabilen** araçlar içerir.
- Veri çözümlemede kullanılabilecek grafiksel araçlara sahiptir.

2-R İSTATİSTİKSEL PROGRAMLAMA DİLİ VERİ DOSYALARI

- Kullanılacak olan veri dosyalarının R ortamına alınabilmesi için farklı seçenekler vardır:
  - metin dosyalarından (txt),
  - hesap tablosu dosyalarından (xls, sav),
  - binary ve dbase (dbf) dosyalarından,
  - gerekli paketleri yükleyerek farklı veritabanlarından da (MySQL, MS Access, Microsoft SQL Server, Postgre SQL, Oracle, IBM DB2) veri almak mümkündür.

2-R İSTATİSTİKSEL PROGRAMLAMA DİLİ KULLANILDIĞI İŞLETİM SİSTEMLERİ

- Açık kaynak kodlu bir yazılım olan R' nin kurulumunun ve kaynak kodunun,
  - Unix,
  - Linux,
  - FreeBSD,
  - Windows,
  - MacOs
- gibi işletim sistemlerinde kullanılabilecek farklı sürümleri R-Project web sitesinden temin edilebilir.

2.1 R'DE NESNE KAVRAMI VE NESNELER

- R, belleğe direkt erişim yerine özel veri yapılarını kullanır. Bu veri yapıları, sembol ve değişkenlerin referans olarak kullanıldığı nesnelerdir.
- R'deki temel nesne türleri şunlardır:
  - **Vektörler**: R' de altı farklı temel vektör tipi bulunmaktadır; logical, integer, real, complex, string ve raw.
  - **Listeler**: Listeler de vektördür ancak listedeki elemanlar farklı tiplerde olabilir.
  - **İfade**: Bir veya daha fazla deyimden oluşan nesnelerdir.
  - **Fonksiyonlar**
  - **NULL**: Özel bir nesnedir. Bir nesnenin boş olup olmadığının belirlenmesi veya boş yapılması için kullanılır.
  - **Ortamlar**: new.env komutu ile oluşturulur. Sembol-değer çiftlerini içeren bir çerçeve ve bir kapsamdan meydana gelir.

2.2 R'DE NESNE OLARAK FONKSİYONLAR

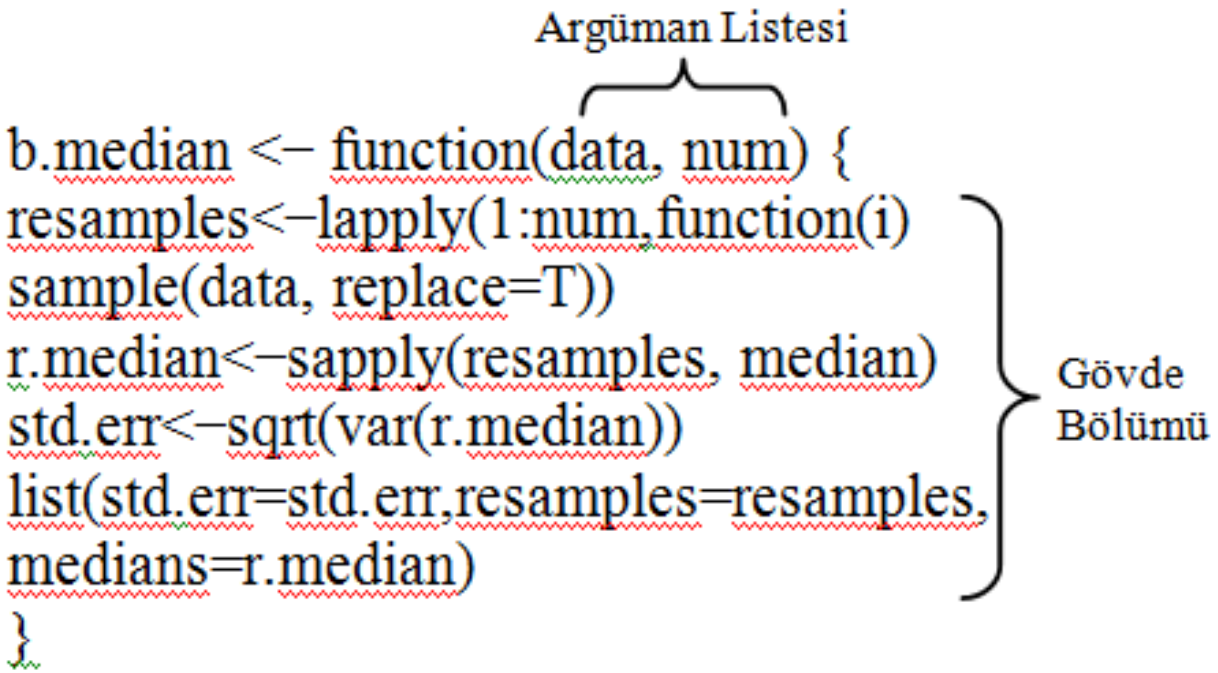
- R'de fonksiyonlar da bir nesne türüdür ve diğer nesneler gibi kullanılır.
- Üç temel bileşeni vardır:
  - Argüman listesi: Bu listede fonksiyonun argümanları virgülle ayrılarak belirtilir.
  - Gövde Bölümü: Tek bir ifade veya değişkenden oluşabildiği gibi bir dizi ifadenin yer aldığı ve "{" ile "}"arasında belirtilen kısımdır.
  - Fonksiyon Ortamı: Fonksiyon oluşturulurken aktif olan ortamdır.

2.2 R'DE NESNE OLARAK FONKSİYONLAR

- R fonksiyonları ayrı paketler halinde düzenlenmişlerdir.
  - Böylece gerekli paketlerle çalışarak daha az bellek kullanımı ve hızlı işlem gücü sağlanır.
  - Bu paketlerin bir başka avantajı da yazılan fonksiyonlardan oluşan paketlerin R web sitesinden temin edilerek yüklenebilmesidir.

- **Install.package()** ve **update.package()** fonksiyonları R komut satırından istenilen paketin indirilmesi ve yüklenmesi için kullanılırlar.

2.2 R'DE NESNE OLARAK FONKSİYONLAR
ÖRNEK FONKSIYON




```
>b1<-b.median(data1,20)
```

- Burada data1 vektörüyle 20 bootstrap örneklem türeterek ortanca ve ortancanın standart hatası tahmini yapılmıştır.

2.3 R KAYNAK KOD(SOURCE CODE) DOSYASI

- R dilinde komut satırına girilen söz dizim kuralları aynı zamanda metin dosyalarına da yazılabilir.
- Bu durumda metin dosyası uzantısı "*.R" olarak kaydedilir. Bu şekilde kaydedilmiş bir dosya artık R script dosyasıdır.
- Bu scriptleri kaynak olarak kullanabilmek için "file-open script" seçeneği kullanılır.

2.3 R KAYNAK KOD(SOURCE CODE) DOSYASI

- Başka bir metin editöründe bulunan komutları tekrar komut satırına yazmadan çalıştırmak da mümkündür.

- File menüsünden "New Script" komutu seçildiğinde açılan R Editor'e istenilen komutlar yazılabilir veya kopyalanabilir. Bu editördeki istenilen satırları çalıştırmak için bu satırları işaretledikten sonra Ctrl+R tuş kombinasyonu kullanılır.

2.4 ÇÖP TOPLAYICI(GARBAGE COLLECTOR)

- R dilinde 1.2.0 sürümünden bu yana kuşak yaklaşımının kullanıldığı bir çöp toplayıcı mevcuttur.
- Çöp toplayıcı programın yazılması ve çalışması sırasında bellek yönetimini gerçekleştirerek olası bellek yönetimi hatalarını önler.
- R' de bellek kullanım durumunu izlemek için "gc" ve "gcinfo" fonksiyonları kullanılır.

2.5 YARDIM KULLANIMI

- R istatistiksel yazılım geliştirme ortamında üç farklı kaynaktan yardım alma imkanı vardır.
  - Çevrim içi yardım,
  - R'nin yardım menüsü,
  - Üçüncüsü ise R-Project web sitesinde bulunan kılavuzlardır.

2.5 YARDIM KULLANIMI

- R dilinin “help” menüsünde bir fonksiyonun nasıl kullanıldığı ve parametrelerinin ne olduğu hakkında yardım sağlayan “fonksiyonel yardım” mevcuttur.
- Fonksiyonel yardıma “Help” menüsünden “R functions” tıklanarak veya komut satırına “help(fonksiyon ismi)” yazarak erişilebilir.
- “help()” fonksiyonu komut satırında iki türlü kullanılabilir:
  - >help(fonksiyon ismi)

  veya
  - >?fonksiyon ismi

2.5 YARDIM KULLANIMI

- Fonksiyon isminin bilinmediği durumlarda ise aşağıdaki üç yöntemle yardım kullanılabilir:
  - help.search('….'):Parantez içinde belirtilenle ilgili yardım konularını listeler. Örneğin; >help.search('data input')
  - find('….'):Parantezde belirtilen kelimenin geçtiği paketin ismini bulur.

    > find('lowess')
  - apropos('….'): Parantez içinde belirtilenle ilgili bütün nesnelerin isimleri bir vektör olarak listelenir.

    > apropos('lm')

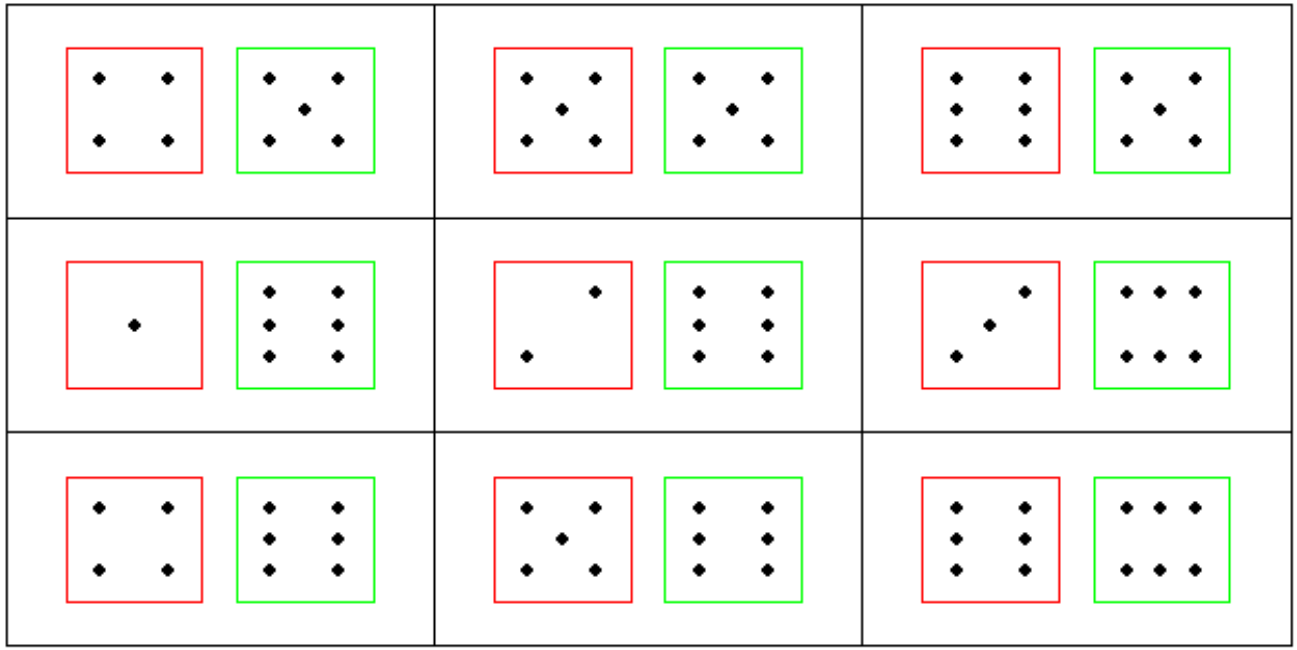
3. SONUÇ

- Bu çalışmada, son yıllarda akademik çalışmalarda yaygın olarak kullanılan R programlama dilinin tanıtılması hedeflenmiştir.
- R, ücretsiz olarak temin edilmesi ve dünyanın çeşitli bölgelerindeki araştırmacıların bu dilin gelişimine destek vermesi sonucunda, özellikle veri işleme ve çözümleme alanlarında çalışan kullanıcıların dikkatini çekmiştir

3. SONUÇ
GÜÇLÜ YÖNLERİ

- ücretsiz temin edilebilmesi,
- nesne yönelimli bir programlama dili olması,
- farklı amaçlar için geliştirilmiş paketler eklenerek fonksiyonelliğinin arttırılabilmesi,
- 2-D, 3-D gelişmiş grafik araçlarına sahip olması,

3. SONUÇ
GRAFİK ÖRNEĞİ



- Hilesiz iki zarın rasgele dokuz kez atılmasını gösteren grafik

```
> library(TeachingDemos)
> plot.dice(expand.grid(1:6,1:6), layout=c(3,3))
```

3. SONUÇ
ZAYIF YÖNLERİ

- Öğrenmesi zor bir programlama dilidir.
- Gelişmiş veri işleme özelliklerine sahip olmasına rağmen bunların kullanılabilmesi özellikle dizi ve matris işlemlerine hâkim olmayı gerektirir.
- Çok büyük veri dosyaları ile çalışmak için uygun değildir. Birkaç yüz megabyte' dan daha büyük veri dosyaları açılmak istendiğinde yetersiz bellek sorunu meydana gelebilir.
- Ticari bir ürün olmadığı için kullanımında karşılaşılan sorunların iletileceği müşteri destek birimi yoktur.

3. SONUÇ

- SAS, SPSS gibi programlar ve R arasındaki en önemli fark şudur: R bir istatistiksel paket program değil istatistiksel yazılım geliştirme ortamı ve programlama dilidir.
- **SAS**: Wegman ve Solka'ya göre istatistik paket programlarının Microsoft'u olarak nitelendirilir. Özellikle veri madenciliği ve bir çok alanda kullanılabilen uygulama araçlarına sahip çok kapsamlı bir paket programdır.
- **SPSS**: Dünya ölçeğinde rekabet gücüne sahip bir başka istatistik paket programıdır ve özellikle sosyal ve eğitim bilimleri alanında kullanıcı bulmaktadır.

3. SONUÇ
R VE S-PLUS

- R ve S-Plus kullanıcı arayüzü dışında pek çok açıdan birbirine benzemektedir. Ancak önemli farklılıklar da mevcuttur:
  - Söz dizimi kurallarındaki farklılıklardan dolayı komutların işletilmesi sonucu farklı sonuçlar çıkabilir.
  - S-Plus verileri diskteki bir dosyada saklar. Bu sayede yaşanılan bir sorun sonucu ortam kaybolmaz. R ise dahili olarak kullanır ve programda bir sorun çıkarsa ortam kurtarılamaz.
  - Bu iki dil arasındaki en önemli fark ise R'nin açık kaynak kodlu olmasıdır.

3. SONUÇ

- Yeni geliştirilen istatistiksel yöntemler için yazılan paketlerin kullanıcılar tarafından kolaylıkla yüklenebilmesi
- İstatistikte önemli bir çalışma alanı olan dayanıklı (robust) istatistiksel yöntemleri kullanmak için gereken fonksiyonlara sahip olması
- Açık kaynak kodlu ve ücretsiz bir programlama dili olması

gibi özellikleri nedeniyle R, akademik çalışmaların yanı sıra istatistik ve matematik eğitiminde de lisanslama problemi olmadan ihtiyaçları karşılayabilecek güçlü bir alternatif oluşturmaktadır

KAYNAKLAR

[1] Braun W.J.,Murdoch D.J.,"A first course in statistical programming", **Cambridge University Press**, England, 1:13-175(2007)

[2] Crawley M. J. , "The R Book", **Wiley serisi**, England, 9-97(2007)

[3] Dalgaard P., "Introductory Statistics with R",**Springer Series**, Denmark, 9-11(2008)

[4] Everitt B. S. , Hothorn Torsten , "A Handbook of Statistical Analyses Using R", London 1-3, 4-5(2005)

[5] http://www.ats.ucla.edu/stat/R/library/bo otstrap.htm, "R Library: Introduction to Bootstrapping", **Ucla Academik Technology Services**

[6]http://www.r-project.org/, "What is R?",(2009)

[7] Ihaka, R., & Gentleman, R., "R: A Language for Data Analysis and Graphics", **Journal of Computational and Graphical Statistics**, 5(3), 299-314 (1996)

[8] Lumley Thomas , "R Fundamentals and Programming Techniques", Birmingham 3-4(2006)

[9] R Core Team, "R Language Definition", 2-8, 26(2008)

[10] R Development Core Team, "R Data Import/Export", 2-4(2008)

[11] R Development Core Team, "R Internals", 11-12(2008)

[12] R Development Core Team, "R: A Language and Environment for Statistical Computing",182-183(2009)

[13] Statistical Computing Group, "Very Basics of R(Windows)", **Research Data Services, University of Pennsylvania** 2-3(2008)

[14] Venables W. N. , Smith D. M. , the R Development Core Team, "An Introduction to R", 2-6(2008)

[15] Wegman E. J. , Solka J. L., "Statistical Software for Today and Tomorrow", in Encyclopedia of Statistics, John Wiley, (2005)