PEMANFAATAN SOFTWARE R UNTUK ANALISIS REGRESI LINEAR

Anita Andriani
D3 Manajemen Informatika, Fakultas Teknologi Informasi
Universitas Hasyim Asy'ari Tebuireng Jombang
Email: anita.unhasy@gmail.com

Abstrak

Software R merupakan alat analisis data yang masuk dalam kategori *freeware*. Software R masih kalah populer jika dibanding dengan software lain seperti SAS, MINITAB, SPSS atau Eviews. Salah satu alasan user lebih memilih software komersil tersebut adalah masih terbatasnya referensi mengenai bahasa dalam pemrograman R. Padahal software R juga mampu memberikan analisis yang kuat dan efektif serta memiliki sistem grafik yang menarik. Salah satu metode dalam statistika yang dapat dikerjakan menggunakan R adalah analisis regresi linear. Pada artikel ini akan dibahas tentang cara memodelkan regresi linear berganda dengan software R sebagai pengganti SPSS. Setelah model regresi linear akan terbentuk dengan bantuan R selanjutnya untuk melihat *fitting* model tersebut dilakukan analisa plot residual.

Kata kunci: Software R, *freeware*, analisis regresi linear

Abstract

Software R is a data analysis tool that falls into the category *freeware*. Software R is still less popular when compared to other software such as SAS, MINITAB, SPSS or Eviews. One of the reasons users prefer this commercial software is that there are still limited references to the language in R programming. Even though R software is also able to provide strong and effective analysis and has an attractive graphics system. One of the methods in statistics that can be done using R is linear regression analysis. In this article, we will discuss how to model multiple linear regression with R software as a substitute for SPSS. After the linear regression model will be formed with the help of R then to see the *fitting of the* model, a residual plot analysis is carried out.

Keywords: Software R, *freeware*, linear regression analysis

I. PENDAHULUAN

R adalah bahasa pemrograman yang digunakan sebagai alat olah data dan komputasi dan bersifat open source. Awalnya R dibuat oleh Robert Gentleman dan Ross Ihaka tahun 1992 di Universitas Auckland yang kemudian dikembangkan oleh R *Development Core Team* dan tersedia secara gratis. Software R dapat dijalankan dalam berbagai sistem operasi seperti Mac, Windows dan Linux.

Salah satu kelebihan dari R adalah *update software* yang relatif cepat dan memiliki kualitas relatif baik karena dikembangkan langsung oleh tim yang ahli dalam bidang statistika. Selain itu, user R di seluruh penjuru dunia turut serta memberikan pengembangan coding, melaporkan adanya bug dan membuat dokumentasi untuk R. Namun sayangnya fasilitas *Graphical User Interface (GUI)* pada R masih kurang begitu memadai karena bersifat *Command Line Interface*. Kemampuan R sebagian besar berasal dari *add-on packages*, yakni kumpulan perintah yang digunakan untuk melakukan suatu analisis. *Package* dasar yang sudah disediakan saat

pertama menginstal R adalah *stats*, *graphics*, *datasets*, *utils*, dan *base*. *Package* lainnya diperoleh dengan cara mengunduh secara online lewat *toolbar install package* atau secara offline dengan mengunjungi website http://cran.rproject.org/.

R dikenal sebagai software statistik karena dapat mengolah dan menganalisa data menggunakan metode atau teknik statistika dengan baik. Tulisan ini membahas penggunaan software R sebagai alternative software statistika lainnya dalam menganalisa model regresi linear. Analisa regresi merupakan teknik statistika yang sering digunakan dalam berbagai bidang. Dengan menggunakan R, kesesuaian dan asumsi model serta multikolinearitas dalam analisis regresi dapat mudah dideteksi. Penulisan artikel dibagi dalam 3 bagian, yaitu latar belakang penulisan, review singkat tentang R dan regresi linear, serta pemodelan regresi linear menggunaan R-CLI yang disertai contoh studi kasusnya. Analisis kecocokan model dibahas menggunakan plot residu.

II. METODE

Pada bagian ini diberikan beberapa dasar software R yang digunakan dalam analisa data beserta konsep regresi linear berganda secara singkat.

2.1 Dasar – Dasar Software R

Subbab ini membahas tentang penggunaan R dalam pengaturan data menggunakan R Commander, meliputi data entry dan import data, jenis – jenis data yang diinput, serta plot data.

2.1.1 Input Data

Entry data R dilakukan dengan cara memasukkan data langsung pada softaware R atau memasukkan data dari sumber lain seperti file Excel, SPSS, text, dll. Perintah untuk melakukan cara yang pertama adalah sebagai berikut:

```
> data1=c(1,2,3,4,5)
> data1
[1] 1 2 3 4 5
```

Misalkan ingin memasukkan sebuah data dengan format teks dengan nama *data.txt*, dan data tersebut disimpan di drive E, maka digunakan perintah read.table()

```
> data2=read.table("E:\\data.txt", header=T)
> data2
   Hari pengunjung pembeli  nominal
1     1         41      27 0.621050
2     2         55      33 1.368500
3     3         39      20 0.571505
4     4         48      23 0.350050
5     5         41      28 0.401100
6     6         27      19 0.273300
7     7         21      19 0.656400
8     8         28      19 1.095350
9     9         32      22 0.541250
10   10         37      25 0.480000
11   11         27      18 0.333550
12   12         30      22 0.711300
13   13         28      21 1.032590
14   14         45      31 0.686400
15   15         36      22 1.099825
```

2.1.2 Jenis Data Objek

Pada pemrograman R, data adalah objek yang memiliki atribut sesuai dengan tipe data seperti: vector, data frame, matriks, dan mode data seperti: *numeric*, *logical*, *complex* dan *character*. Agar data berbentuk vektor maka dapat menggunakan fungsi `c()` seperti contoh berikut:

```
> x=c(0,7,8)
> x
[1] 0 7 8
```

Seandainya ingin dibuat matriks yang banyak baris dan kolomnya masing - masing adalah 2 dan 3, maka digunakan Fungsi `matrix ()`

```
> m=matrix(1:6, nrow=2, ncol=3)
> m
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
```

Untuk menggabungan kolom atau baris yang baru ke sebuah matriks dapat dilakukan dengan menuliskan perintah `rbind` jika ingin menambah baris, `cbind` jika kolom yang ingin ditambahkan.

```
> m2=cbind(m,c(7,8))
> m2
     [,1] [,2] [,3] [,4]
[1,]    1    3    5    7
[2,]    2    4    6    8
> m3=rbind(m,c(7,8,9))
> m3
     [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
[3,]    7    8    9
```

Sebuah objek dengan mode data yang berbeda dapat digabung dalam satu data dengan menggunakan data frame.

```
>     dataframe=data.frame(nomer=1:4,nama=c('ani','ita','andri','nita'),
nilai=7:10)
> dataframe
  nomer  nama nilai
1     1   ani     7
2     2   ita     8
3     3 andri     9
4     4  nita    10
```

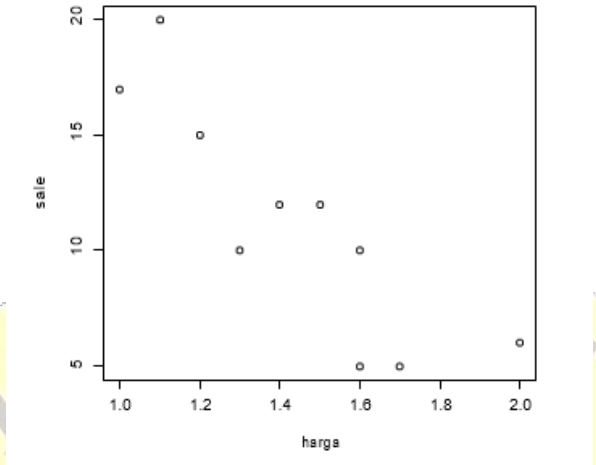
2.1.3 Plot data

Grafik pada R ditampilkan menggunakan fungsi plot yang terbagi menjadi 3 jenis fungsi, yaitu:

1. Plot utama, membuat plot yang baru di jendela grafik dengan menulis perintah `qqplot,plot, image, hist,persp, dan contour.`
2. Plot tambahan, memberikan informasi tambahan ke gambar yang telah ada sebelumnya dengan bantuan plot utama, misalkan ingin menambahkan titik atau garis baru ataupun keterangan untuk grafik maka dapat menuliskan `lines, points, abline, text, abline,title, atau legend.`
3. Fungsi yang bersifat interaktif, menambah atau mengumpulkan informasi berdasarkan plot yang tersedia dengan menulis perintah `locator, identify.`

Contoh dalam menggunakan fungsi plot akan disajikan sebagai berikut:

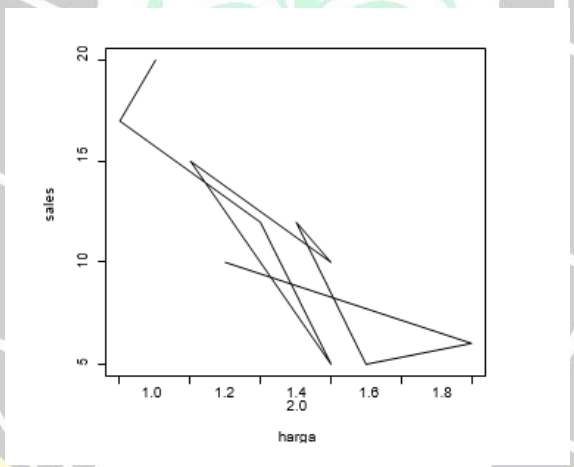
```
> harga=c(1.3,2,1.7,1.5,1.6,1.2,1.6,1.4,1,1.1)
> sales=c(10,6,5,12,10,15,5,12,17,20)
> plot(harga,sales)
```



Gambar 1. Plot Penjualan

Jika grafik pada Gambar 1 diberi fungsi tambahan **lines,** maka grafiknya akan menjadi:

```
> plot(harga,sales, type='l')
```



Gambar 2. Plot Penjualan dengan Fungsi Tambahan Lines

2.2 REGRESI LINEAR

Regresi adalah salah satu teknik statistik yang menjelaskan pola hubungan antar dua variabel atau lebih. Variabel pada regresi dibedakan menjadi dua, diantaranya adalah:

I. Variabel respon (variabel dependent), merupakan variabel tidak bebas yang dipengaruhi oleh variabel lainnya dan dinotasikan dengan Y.
II. Variabel prediktor (variabel independent), merupakan variabel bebas yang tidak dipengaruhi oleh variabel lainnya dan dinotasikan dengan X

Analisa regresi memberikan informasi mengenai ada tidaknya hubungan, pengaruh dan besarnya hubungan antar variabel penelitian.

Jika variabel bebas banyaknya lebih dari satu maka disebut dengan regresi linear berganda. Analisis regresi linear berganda bertujuan untuk mengetahui tingkat hubungan antara variabel dan membuat perkiraan nilai Y terhadap X. Secara umum regresi linear dimodelkan sebagai berikut:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Dengan β_0, β_1 adalah koefisien regresi. Taksiran dari model tersebut berdasarkan n random sampel sebagai berikut:

$$\hat{Y} = b_0 + b_1 X_1$$

dengan

$\hat{Y}$ = nilai estimasi variabel Y

b_0 = nilai estimasi parameter β_0

b_1 = nilai estimasi parameter $\beta_1, \ldots, \beta_n$

Nilai b_0 dan b_1 dicari dengan menggunakan metode least square, secara sederhana dituliskan sebagai berikut:

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i{}^2) - (\sum_{i=1}^{n} x_i)^2}$$

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i{}^2}{n}$$

III. HASIL DAN PEMBAHASAN

Bagian ini akan menyajikan hasil analisis data dan plot residu pada studi kasus yang dipilih menggunakan versi R-CLI.

3.1 Analisa Data

Diberikan sebuah data sebagai berikut:

Tabel 1. Data Transaksi Pembelian di Indomart Kedung Mundu Semarang

$Hari\ Ke-$	$Pengunjung\ (X_1)$	$Pembeli\ (X_2)$	$Nominal\ Pembelian\ (dlm\ jutaan)\ (Y)$
1	41	27	0.62105
2	55	33	1.3685
3	39	20	0.57151
4	48	23	0.35005
5	41	28	0.4011
6	27	19	0.2733
7	21	19	0.6564
8	28	19	1.09535
9	32	22	0.54125
10	37	25	0.48
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
24	42	27	0.53925
25	45	34	0.481
26	35	23	0.34285
27	45	35	0.70285
28	51	31	0.76075
29	48	33	0.64903
30	40	24	1.05513

Berdasarkan data pada Tabel 1 ingin diketahui apakah jumlah kedatangan pengunjung dan kedatangan pembeli di indomart Kedung Mundu Semarang mempengaruhi nomimal pembeliannya. Langkah pertama yang dilakukan dalam analisis regresi linear berganda dengan R adalah memanggil data yang telah disimpan di direktori dengan cara:

```
> data=read.table("E:\\data.txt", header=T)
```

Persamaan regresi linear berganda dengan metode least squared diperoleh dengan memanggil fungsi **lm**:

```
> hasilanalisis=lm(nominal~pengunjung+pembeli,data)
```

Hasil estimasi koefisien – koefisien ini dapat dilihat dengan menuliskan

```
> hasilanalisis
```

maka akan muncul keterangan seperti:

```
Call:
lm(formula = nominal ~ pengunjung + pembeli, data = data)

Coefficients:
(Intercept)   pengunjung      pembeli
   0.458728     0.006453    -0.003463
```

Dari hasil analisa tersebut terbentuklah persamaan regresi berikut ini:

$$Nominal = 0.458728 + 0.006453 pengunjung - 0.003463 pembeli \quad \ldots (1)$$

Koefisien – koefisien regresi dan beberapa statistic lainnya dapat pula dilihat dengan menggunakan fungsi

```
summary:
Call:
lm(formula = nominal ~ pengunjung + pembeli, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33877 -0.17593 -0.07371  0.07895  0.66914

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.458728   0.262020   1.751   0.0913 .
pengunjung   0.006453   0.010466   0.617   0.5427
pembeli     -0.003463   0.016267  -0.213   0.8330
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2741 on 27 degrees of freedom
Multiple R-squared:  0.02215,   Adjusted R-squared:  -0.05028
F-statistic: 0.3058 on 2 and 27 DF,  p-value: 0.7391
```

Berdasarkan tampilan pada summary, intercept dalam hal ini adalah konstanta serta koefisien kedua variabel, memiliki nilai yang sama seperti pada persamaan (1). Dilihat dari Pr(> | t |) dari variabel pengunjung, Pr(>| t |) = 0,5427 > 0,05 artinya variabel pengunjung tidak signifikan atau tidak berpengaruh terhadap nominal pembelian. Hal yang sama juga terjadi pada variabel pembeli dengan nilai Pr(> | t |) = 0.8330 > 0,05, dengan demikian variabel pembeli juga tidak terlalu berpengaruh terhadap nominal pembelian di indomart Kedung Mundu Semarang. Secara teori kedua variabel yang tidak signifikan ini harus dikeluarkan dari model. Akan tetapi tidak menutup kemungkinan bahwa variabel yang tidak signifikan tersebut tetap dipertahankan seperti pada kasus ini. Berdasarkan nilai R-squared sebesar 0,02215, artinya bahwa penelitian ini hanya mampu menjelaskan 2,215% keragaman nominal pembelian ditentukan oleh banyaknya pembeli dan pengunjung, selebihnya ditentukan oleh faktor lain.

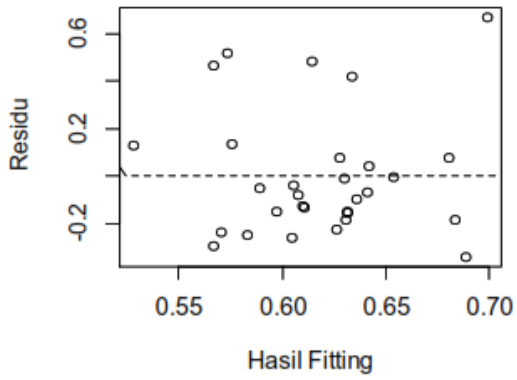
3.2 Analisis Plot

Residu adalah perbedaan nilai data observasi dengan nilai estimasi model. Biasanya residu ditampilkan dalam bentuk diagram atau plot, pada artikel ini akan ditampilkan plot antara residu dengan nilai estimasi dan plot probabilitas normal dari residu.

3.2.1 Plot Residu dengan Nilai Estimasi

Plot antara residu dengan nilai estimasinya ditampilkan dengan perintah:

```
> residu=residuals(hasilanalisis)
> fitting=fitted(hasilanalisis)
> plot(fitting,residu, xlab="Hasil Fitting", ylab="Residu")
> abline(h=0, lty=2)
> text(fitting,residu,labels=rownames(hasilanalisis))
```



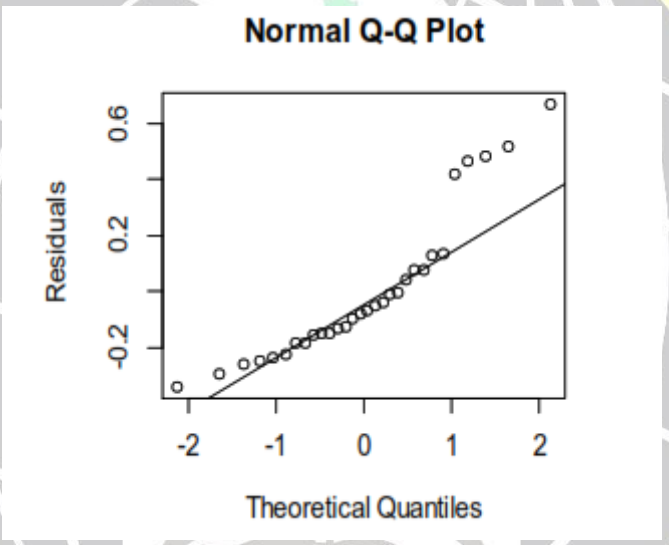
Gambar 3. Plot Residu VS Hasil Estimasi

Gambar 3 memperlihatkan bahwa tidak ada pola yang jelas yang dapat ditunjukkan dari plot tersebut. Sehingga disimpulkan bahwa plot residu dengan hasil estimasi menunjukkan tidak ada masalah terhadap model.

3.2.2 Plot Probabilitas Normal Residu

Plot probabilitas normal residu dapat ditampilkan dengan perintah sebagai berikut:

```
> qqnorm(residu, ylab="Residuals")
> qqline(residu)
```



Gambar 4. Plot Probabilitas Normal

Gambar 4 memperlihatkan bahwa ada beberapa sampel yang menjauhi garis linear. Banyaknya sampel yang berada di garis linea menunjukkan bahwa tidak ada indikasi ketidaknormalan data pada residu.

IV. KESIMPULAN DAN SARAN

Berdasarkan hasil analisa sebelumnya, maka model regresi yang dihasilkan adalah:

$$Nominal = 0.458728 + 0.006453 pengunjung - 0.003463 pembeli$$

dimana kedua variabel pengunjung dan pembeli tidak berpengaruh secara signifikan secara statistik. Akan tetapi kedua variabel tetap dipertahankan karena keterbatasan variabel penelitian. Berdasarkan nilai R-squared sebesar 0,02215, kedua variabel juga hanya mampu menjelaskan 2,215% keragaman nominal pembelian. Namun jika dilihat dari plot antara residu dengan nilai estimasi dan plot probabilitas normal dari residu, model masih bisa diterima.

V. DAFTAR PUSTAKA

Brian S. Everitt and Torsten Hothorn. 2007. *A Handbook of Statistical Analyses Using R*. A Chapman & Hall Book.

Paradis, Emmanuel. 2005. *R for Beginners*. France: Institut des Sciences de l'Evolution.

Pratomo, Dedi Suwarsito dan Erna Zuni Astuti. 2015. *Analisis Regresi dan Korelasi Antara Pengunjung dan Pembeli Terhadap Nominal Pembelian di Indomaret Kedungmundu Semarang Dengan Metode Kuadrat Terkecil*. E-print Jurnal Udinus.

Walpole, Myers, and Ye. 2012. *Probability & Statistics for Engineers & Scientists Ninth Editio*n. Prentice Hall.

Venables, W.N, D. M. Smith and the R Core Team. 2017. *An Introduction to R, Notes on R: A Programming Environment for Data Analysis and Graphics Version 3.4.2 (2017)*. R Project.