

# **APLIKASI SOFTWARE R DALAM ANALISIS REGRESI**

**DRS. OPEN DARNIUS M.Sc**

**Fakultas Matematika dan Ilmu Pengetahuan Alam  
Jurusan Matematika  
Universitas Sumatera Utara**

## **1. Pendahuluan**

R adalah suatu software terintegrasi yang memiliki fasilitas untuk pemanipulasian data, perhitungan, dan penampilan grafik. Software ini terus dikembangkan dan diberikan secara gratis kepada masyarakat yang memerlukannya dengan cara mendownload melalui internet. R dapat dikatakan sebagai suatu implementasi bahasa S yang telah dikembangkan di Bell Laboratories oleh Rick Becker, John Chamber dan Allan Willks, orang yang juga membentuk dasar system S-Plus. Walaupun lingkungan R (R environment) tidak menyebutkan statistik, namun banyak orang menggunakan R sebagai system statistik. Hal ini dikarenakan kemampuan R dalam mengolah dan menganalisis data cukup baik. Oleh karena itu dalam tulisan ini akan disajikan salah satu pemakaian R dalam analisis statistika, yaitu analisis regresi.

Analisis regresi merupakan metode analisis yang sering digunakan dan telah dipakai dalam banyak bidang yang berbeda seperti ekonomi, bisnis, dan pendidikan. Namun analisis ini masih sering digunakan secara keliru, dan pemakaiannya sering hanya melibatkan estimasi kwadrat terkecil (least-squares estimation) dari koefisien regressinya, ditambah sekilas pandang tentang diagram pencar residunya. Secara umum, analisis regresi harus mempertimbangkan dengan cermat beberapa hal seperti: kesesuaian model yang digunakan, asumsi-asumsi model, dan masalah multikolinieritas serta pendeteksian data pencilan. Membahas semua hal di atas dalam menyelesaikan suatu permasalahan analisis regresi memerlukan perhitungan yang rumit, khususnya bila melibatkan beberapa variabel dengan data observasi yang besar. Namun perhitungan ini bukanlah menjadi masalah yang besar jika dikerjakan dengan alat komputer. Software R sebagai salah satu alat komputasi dapat digunakan untuk analisis ini, oleh karenanya tulisan ini diberi judul Aplikasi Software R dalam Analisis Regresi.

Tulisan ini dengan singkat menyajikan secara teoritis tentang analisis regresi. Selanjutnya suatu contoh kasus yang dikerjakan dengan menggunakan software R akan diberikan sebagai bahan diskusi. Pemakaian software R secara eksplisit akan disajikan dalam studi kasus, dan analisis hasil (output) nya akan dijelaskan. Analisis ketepatan model regresi dalam tulisan ini dibahas hanya dengan analisis plot dari residu.

## **2. Analisis Regresi**

Regresi adalah suatu alat untuk mengeksplorasi hubungan antar peubah. Masalah regresi pada umumnya adalah masalah yang berkaitan dengan pengestimasiian atau pendugaan nilai dari suatu peubah acak terikat Y (dependent variable) berdasarkan satu atau beberapa ukuran yang diketahui dari satu atau beberapa peubah bebas X (independent variable(s)). Jika suatu peubah acak terikat Y diduga berdasarkan hanya satu peubah bebas X, maka analisisnya disebut analisis regresi sederhana. Tapi jika suatu peubah acak terikat Y diduga berdasarkan dua atau lebih peubah bebas X, maka analisisnya disebut analisis regresi ganda. Sebagai contoh ingin diduga besarnya hasil penjualan (Y) dari suatu produk tertentu

berdasarkan besarnya biaya advertensi/ promosi (X), merupakan kasus regresi sederhana. Namun bila ingin diduga jumlah hasil panen padi (Y) berdasarkan jumlah pemakaian pupuk (X1), jumlah curah hujan (X2), dan luas areal pertanian (X3), maka disebut kasus regresi ganda.

### 2.1. Model umum Regresi linier

Jika Y adalah peubah acak terikat (tak bebas) dan  $X_j$  adalah k-peubah acak bebas, dimana  $j=1,2,\dots,k$ , maka model regresi linier yang lazim digunakan adalah:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \varepsilon_i, \text{ untuk } i = 1, 2, \dots, n$$

(1)

dimana  $\varepsilon_i$  = error pada observasi ke i.

Model ini dapat ditulis dalam model bentuk matrix sebagai berikut:

$$\mathbf{Y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

(2)

Dimana

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \text{dan} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

### 2.2. Asumsi Standard

Ada lima asumsi standard (baku) yang secara umum diberikan pada model (1) dalam memberdayakan pemakaian metode kwadrat terkecil, yaitu:

1. Baik untuk  $x_{1i}, x_{2i}, \dots, x_{ki}$  yang tetap ataupun sebagai suatu realisasi dari peubah acak  $X_{1i}, X_{2i}, \dots, X_{ki}$ , adalah bebas dari suku error ( $\varepsilon_i$ )
2. Suku error  $\varepsilon_i$  adalah peubah acak dengan rata-rata nol, yaitu  $E[\varepsilon_i] = 0$ ,
3. Peubah acak  $\varepsilon_i$  mempunyai varians yang sama, yaitu  $E[\varepsilon_i^2] = \sigma_\varepsilon^2$ .
4. Peubah acak  $\varepsilon_i$  tidak berkorelasi satu dengan lainnya, yaitu  $E[\varepsilon_i \varepsilon_j] = 0$ , untuk  $i \neq j$ .
5.  $X_{1i}, X_{2i}, \dots, X_{ki}$  adalah bebas linier.

### 2.3. Metode Kwadrat Terkecil

Metode Kwadrat Terkecil (Least Square Method) merupakan suatu metode dalam pemilihan garis regresi linier yang meminimumkan jumlah kwadrat residu, yaitu jumlah kwadrat dari selisih nilai observasi peubah responsnya dengan nilai dugaan. Metode ini dapat dijelaskan sebagai berikut:

Andaikan dimiliki suatu model seperti (2). Konsep dari metode kwadrat terkecil adalah menentukan  $\boldsymbol{\beta}$ , yang meminimumkan jumlah kwadrat residu  $\boldsymbol{\varepsilon}^t \boldsymbol{\varepsilon}$ , dimana superscrip t menyatakan matriks transpose. Jumlah kwadrat residu ini dapat dituliskan dalam bentuk fungsi  $\boldsymbol{\beta}$ , yaitu:

$$\Psi(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^t(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \tag{3}$$

Karena  $\Psi(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^t(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$  merupakan suatu fungsi kwadrat non-negatif bernilai riil maka keberadaan minimum berhingga dari  $\Psi(\boldsymbol{\beta})$  dapat dijamin. Penyelesaian minimisasi fungsi ini dapat dilakukan dengan menyelesaikan persamaan (3) menjadi persamaan berikut:

$\boldsymbol{\Psi}(\boldsymbol{\beta}) = \mathbf{Y}^t \mathbf{Y} - \boldsymbol{\beta}^t \mathbf{X}^t \mathbf{X} \boldsymbol{\beta} - 2\boldsymbol{\beta}^t \mathbf{X}^t \mathbf{Y}$ , dan dengan mendiferensialkan secara parsial persamaan ini terhadap  $\boldsymbol{\beta}$ , diperoleh:

$\delta \boldsymbol{\Psi}(\boldsymbol{\beta}) = (\mathbf{X}^t \mathbf{X} \boldsymbol{\beta} - \mathbf{X}^t \mathbf{Y}) \delta(\boldsymbol{\beta})$ . Dengan menyamakan dervatif ini sama dengan nol, maka penyelesaian  $\boldsymbol{\beta}$ , dinotasikan dengan  $\hat{\boldsymbol{\beta}}$  yang meminimumkan fungsi  $\boldsymbol{\Psi}(\boldsymbol{\beta})$  diperoleh sebagai berikut:

$$\mathbf{X}^t \mathbf{X} \beta = \mathbf{X}^t \mathbf{Y} \quad (4)$$

Persamaan ini dikenal dengan persamaan normal. Penduga (estimator)  $\beta^*$ , selanjutnya dapat diperoleh dari (4) sebagai berikut:

$$\beta^* = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Y}$$

Penduga  $\beta^*$ , merupakan penduga tak bias (unbiased estimator), dimana  $\mathbf{E}(\beta^*) = \beta$ , dengan kovarians  $\sigma(\mathbf{X}^t \mathbf{X})^{-1}$ . Dan model persamaan garis regresi liniernya adalah:

$$\mathbf{Y}^* = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Y} \quad (5)$$

### 3. Suatu Contoh Kasus

Dalam bidang pendidikan diduga bahwa keberhasilan seorang mahasiswa dalam suatu matakuliah bergantung kepada nilai ujian saringan masuk, dan frekwensi ketidakhadirannya mengikuti matakuliah tersebut. Untuk keperluan ini secara acak diambil 30 mahasiswa yang mengikuti matakuliah statistika dasar, dan dicatat nilai statistiknya (Y), nilai ujian saringan masuknya ( $X_1$ ), dan frekwensi ketidakhariannya ( $X_2$ ) sebagai berikut:

No	Y	$X_1$	$X_2$
1	85	65	1
2	74	50	7
3	76	55	5
4	90	65	2
5	85	55	6
6	87	70	3
7	94	65	2
8	98	70	5
9	81	55	4
10	91	70	3
11	75	50	1
12	74	55	4
13	90	65	2
14	85	55	6
15	87	70	3

No	Y	$X_1$	$X_2$
16	94	65	2
17	98	70	5
18	81	55	4
19	94	65	2
20	98	70	5
21	81	55	4
22	91	70	3
23	75	50	1
24	74	55	4
25	70	50	4
26	80	60	3
27	88	64	1
28	93	74	3
29	76	55	2
30	83	60	3

Selanjutnya akan diramalkan berapa nilai statistika dasar mahasiswa yang diketahui nilai ujian saringan masuknya, dan frekwensi ketidakhadirannya. Untuk keperluan ini akan ditentukan model regresi liniernya dengan menggunakan metode kwadrat terkecil. Selanjutnya model ini akan dianalisis, dengan menguji asumsi-asumsi yang dikenakan terhadap model tersebut.

### 4. Analisis Dengan R

Andaikan data di atas sudah disimpan dalam file ACII dengan nama **nilai.data**, tanpa adanya label baris maupun kolom. Langkah pertama yang dilakukan dalam analisis dengan R adalah membentuk suatu data frame R. hal ini dapat dilakukan sebagai berikut, setelah software R dieksekusi:

```
>nilai.dat <- read.table ("nilai.data",col.names =
c("n.stat","n.usm","f.absn"));
```

Perintah ini membentuk data frame dalam R dengan nama data frame **nilai.dat**, dimana argumen `col.names = c("n.stat","n.usm","f.absn")` memberikan nama variabel Y menjadi n.stat, variabel  $X_1$  menjadi n.usm, dan variabel  $X_2$  menjadi f.absn dari data nilai.data.

#### 4.1. Menentukan Persamaan Regressi

Untuk mendapatkan persamaan regresi linier dengan metode kuadrat terkecil, dari model  $n.stat = \alpha + \beta_1 n.usm + \beta_2 f.absn + \varepsilon$  dalam R, dapat dilakukan dengan fungsi **lm**. Hal ini dapat dilakukan sebagai berikut:

```
> nilai.fit<-lm(n.stat~n.usm+f.absn, nilai.dat);
```

Perintah ini menghasilkan nilai.fit yang menyimpan nilai estimasi intercept (konstan) dan koefisien n.usm, dan f.absn.

Untuk melihat hasil estimasi koefisien-koefisien ini dapat dilihat dengan menuliskan

```
> nilai.fit, dan enter.
```

Hasilnya akan diperoleh sebagai berikut:

Call:

```
lm(formula = n.stat ~ n.usm + f.absn, data = nilai.dat)
```

Coefficients:

(Intercept)	n.usm	f.absn
23.6432	0.9767	0.4845

#### Analisis

Hasil ini menunjukkan intercept dan koefisien dari kedua variabel n.usm dan f.absn. Sehingga persamaan regresi linier yang diperoleh adalah:

$$n.stat = 23.6432 + 0.9767 n.usm + 0.4845 f.absn \quad (6)$$

Koefisien-koefisien hasil ini dan beberapa statistik lainnya dapat juga dilihat dengan menggunakan fungsi **summary**, hal ini dapat dilakukan dengan menuliskan

```
> summary (nilai.fit); , lalu enter. Hasilnya akan diperoleh sebagai berikut:
```

Call:

```
lm(formula = n.stat ~ n.usm + f.absn, data = nilai.dat)
```

Residuals:

Min	1Q	Median	3Q	Max
-6.464		-3.426	1.533	3.185
				5.904

Coefficients:

Estimate		Std. Error	t value	Pr(> t )
(Intercept)	23.6432	6.7515	3.502	0.00163 **
n.usm	0.9767	0.1020	9.575	3.57e-10 ***
f.absn	0.4845	0.4796	1.010	0.32138

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.075 on 27 degrees of freedom

Multiple R-Squared: 0.7736, Adjusted R-squared: 0.7568

F-statistic: 46.12 on 2 and 27 DF, p-value: 1.957e-09

#### Analisis

Bagian pertama dari hasil analisis dengan fungsi **summary** ini adalah ringkasan lima angka (five number summary) dari residu, yaitu nilai minimum, kuartil pertama, median, kuartil ketiga, dan nilai maximum. Tampilan ini menunjukkan adanya penyebaran residu yang mendekati normal. Pada bagian kedua ditampilkan nilai estimasi dari intercept dan koefisien dari kedua variabel n.usm, dan f.absn disertai dengan standard error, nilai statistik t (t-value) dan Pr(>|t|) masing-masing.

Dari tampilan ini dapat juga dilihat bahwa persamaan regresi liniernya adalah sama dengan (6). Dari nilai statistik t, dan nilai  $Pr(>|t|)$  dari koefisien regresi n.usm  $Pr(>|t|) = 3.57e-10 < 0.05$ , yang berarti koefisien regresi n.usm bersifat nyata, sedangkan koefisien regresi f.absn mempunyai  $Pr(>|t|) = 0.32138 > 0.05$  (taraf nyata). Berarti frekwensi absen tidak berpengaruh nyata terhadap keragaman nilai ujian statistika dasar. Karenaya variabel ini harus dikeluarkan dari model regresi linier ganda. Bagian terakhir hasil ini menunjukkan nilai standard error residu dari model adalah 4.075, nilai R squared adalah 0.7736, yang artinya 77,36% keragaman nilai ujian statistika dasar ditentukan oleh besarnya nilai ujian saringan masuk (n.usm) dan frekwensi absen (f.absn) seorang mahasiswa. Selebihnya 22.64% ditentukan oleh factor lain. Statistik F = 46.12 lebih besar dari nilai  $F_{0.05(2,28)} = 3.34$  dengan peluang (p-value=1.957e-09) lebih kecil dari taraf nyata 0.05. Dengan demikian dapat disimpulkan bahwa hubungan antara variabel n.usm dan f.absn dengan nilai ujian statistika dasar dalam persamaan **n.stat = 23.6432 + 0.9767 n.usm + 0.4845 f.absn** bersifat nyata.

#### 4.2. Analisis Plot dari Residu

Residu ( $r_i$ ) adalah perbedaan antara nilai data pengamatan ( $y_i$ ) dengan nilai estimasi data dari model ( $y_i^*$ ), sehingga ketepatan (keakuratan) dari suatu model dapat dilihat dari residu. Oleh karena itu langkah selanjutnya dilakukan dalam analisis regresi adalah analisis residu, yaitu pengujian terhadap residu dari model yang sudah diperoleh. Secara visual pada umumnya akan ditampilkan beberapa diagram atau plot, dimana plot yang paling berguna dan banyak dipakai dalam analisis ini antara lain adalah:

1. Plot antara residu dengan masing-masing independent variabel. Misalnya adanya hubungan curvilinear, menunjukkan perlu ditambahkan kedalam model bentuk order (pangkat) yang lebih tinggi dari variabel independennya, sebagai contoh penambahan suku kwadrat dari variabel independennya.
2. Plot antara residu dengan nilai estimasi. Jika varians dari residu tampak membesar dengan membesarnya nilai estimasi, maka suatu transformasi nilai observasi seharusnya dilakukan sebelum membentuk persamaan regresi.
3. Plot peluang normal dari residu. Setelah semua variasi yang sistematis dikeluarkan dari data, residu seharusnya kelihatan seperti sampel dari distribusi normal. Suatu plot residu berurut dengan ekspektasi statistik berurut dari distribusi normal memberikan uji asumsi ini.

Sebelum menggambarkan plot ini, nilai nilai residu dan nilai estimasi perlu ditentukan. Hal ini dapat dilakukan dengan R sebagai berikut:

Untuk mendapatkan nilai residu digunakan **fungsi residuals**, untuk kasus di atas cukup ditulis dengan:

```
>nilai.res<-residuals(nilai.fit)
```

Untuk mendapatkan nilai estimasi digunakan **fungsi predict**, untuk kasus di atas cukup ditulis dengan

```
>nilai.prd<-predict(nilai.fit)
```

Dengan memperoleh nilai residu yang tersimpan dalam **nilai.res**, dan nilai estimasi dalam **nilai.prd**, maka pembuatan grafik (plot) residu yang diinginkan sudah dapat dilakukan, hanya saja residu yang diperoleh dalam hal ini belum dalam bentuk yang baku, hingga perlu ditentukan nilai baku residu ini. Dalam R hal ini dapat dilakukan sebagai berikut:

```
>s<- summary(nilai.fit)$sigma
```

```
>h<-lm.influence(nilai.fit)$hat
```

```
>nilai.res<-nilai.res/(s*sqrt(1-h))
```

Perintah pada baris pertama adalah untuk mendapatkan standard deviasi dari nilai estimasi, perintah pada baris kedua adalah untuk mendapatkan nilai

diagonal matrik dari matrik  $\mathbf{H} = \mathbf{X}(\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t$ , dan perintah pada baris ketiga adalah menghitung nilai residu baku.

Penggambaran plot residu sudah dapat dilakukan, hal ini dapat dilakukan sebagai berikut:

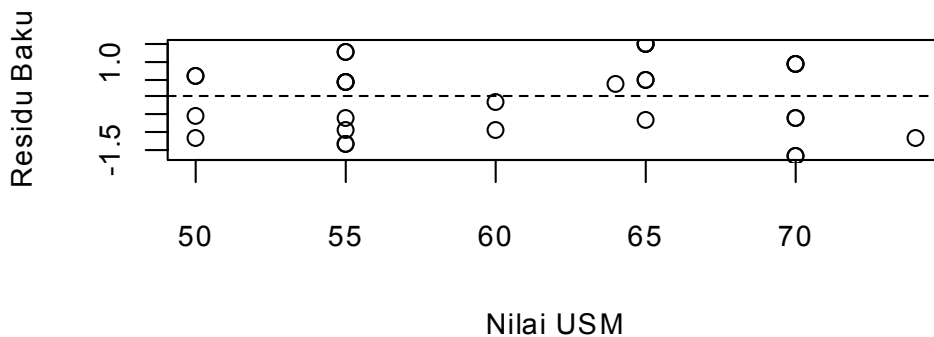
#### 4.2.1. Plot Residu dengan Variabel Terikat

Yang dimaksud dengan residu disini adalah residu baku. Plot ini dapat dilakukan dalam R dengan perintah sebagai berikut:

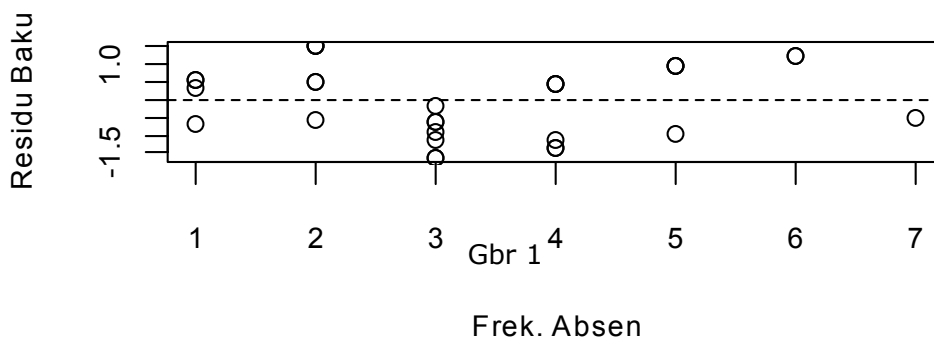
```
>plot(mfrow=c(2,1)
>plot(nilai.dat[,"n.usm"],nilai.res,xlab="Nilai USM",ylab="Residual Baku")
>abline(h=0,lty=2)
>title("Residu Baku Vs Nilai USM")
>plot(nilai.dat[,"f.absn"],nilai.res,xlab="Frek. Absen",ylab="Residual Baku")
>abline(h=0,lty=2)
>title("Residu Baku Vs Nilai USM")
```

Tampilan grafik dari perintah diatas ditunjukkan dalam Gbr. 1 di bawah ini

#### Residu baku Vs Nilai USM



#### Residu baku Vs Frek. Absen



#### Analisis

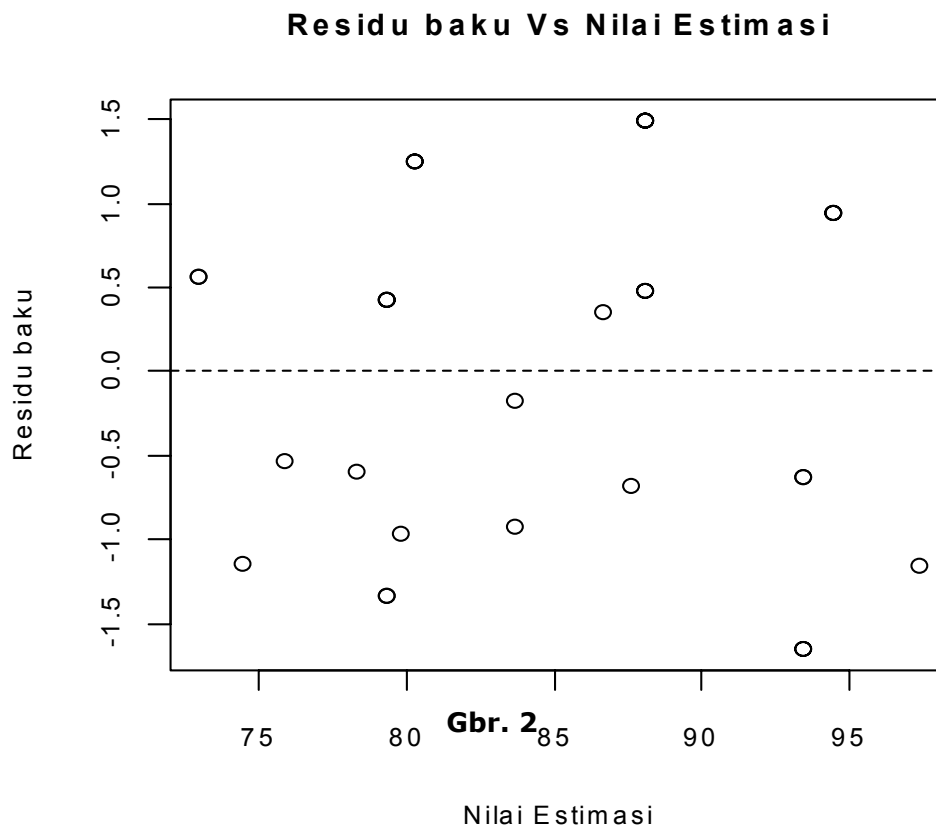
Plot residu baik dengan variabel n.usm maupun f.absn di atas menunjukkan tidak adanya masalah yang mendasar dari model, hal ini ditunjukkan dengan tidak adanya kenaikan/penurunan variance maupun hubungan curvilinear antara residu dengan kedua variabel terikat.

#### 4.2.2. Plot Residu dengan Nilai Estimasi

Untuk menampilkan plot residu baku dengan nilai estimasi dalam R dapat dilakukan dengan perintah berikut:

```
>plot(mfrow=c(1,1))
>plot(nilai.prd,nilai.res,xlab="Nilai Estimasi",ylab="Residual Baku")
>abline(h=0,lty=2)
>title("Residu Baku Vs Nilai Estimasi")
```

Tampilan grafik dari perintah diatas ditunjukkan dalam Gbr. 2 di bawah ini



#### Analisis

Plot Residu vs nilai estimasi ini juga menunjukkan tidak ada masalah yang mendasar terhadap model yang diperoleh, karena tidak ada suatu pola yang jelas dapat ditunjukkan dari plot tersebut.

#### 4.2.3. Plot Peluang Normal Residu

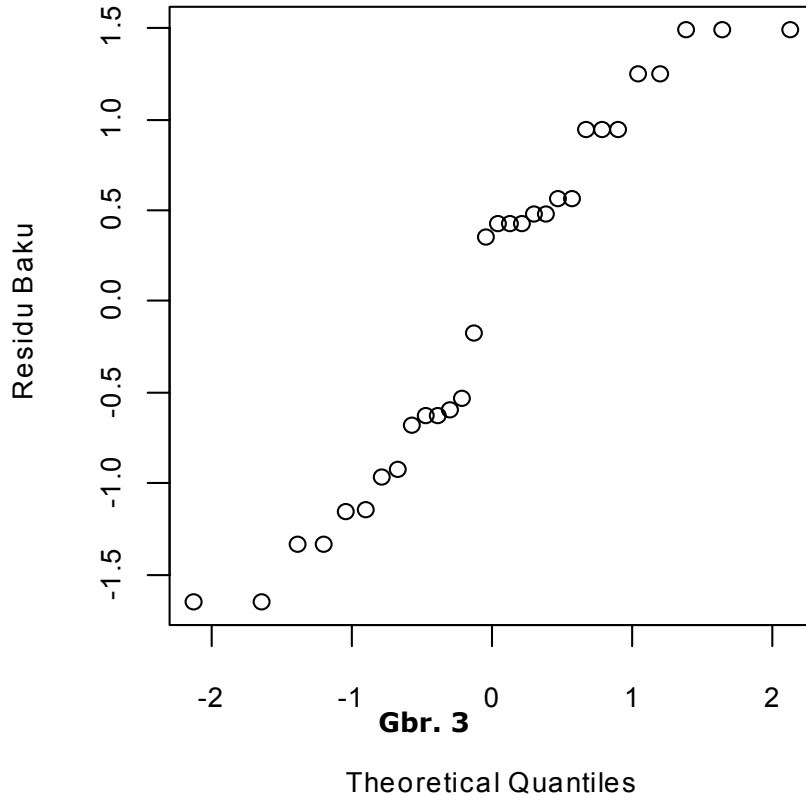
Untuk menampilkan plot residu baku dengan nilai estimasi dalam R dapat dilakukan dengan perintah berikut:

```
>par(pty)
>qqnorm(nilai.res,ylab="Residu Baku")
```

```
>title("Plot Normal Residu Baku")
```

Tampilan grafik dari perintah ini ditunjukkan dalam Gbr. 3 di bawah ini.

### **Plot Normal dan Residu baku**



### **Analisis**

Interpretasi plot ini tidak cukup jelas, terlebih untuk sampel yang kecil. Walaupun demikian kita bisa menyatakan tidak ada indikasi ketidak normalan dari residu. Hal ini kita nyatakan dengan melihat kecenderungan plot yang hampir linier.

### **5. Kesimpulan dan Saran**

Dari analisis dapat dilihat bahwa model regresi linier dengan menggunakan kedua variabel bebas adalah:

$$\mathbf{n.stat = 23.6432 + 0.9767 n.usm + 0.4845 f.absn}$$

dimana sesungguhnya variabel **f.absn** tidak berpengaruh nyata secara statistik, sehingga variabel ini dapat dikeluarkan dari model. Namun persamaan di atas setelah dianalisis tidak akan memberikan kekurang tepatan hasil peramalan. Hal ini sesungguhnya dapat dipahami karena dalam model regresi jika suatu variabel penduga yang independent secara nyata sekalipun dengan variabel penjelas diikuti sertakan dalam model maka tidak akan memberikan pengaruh yang lebih jelek terhadap peramalan.

## **6. Daftar Bacaan**

- [1] Brian S. Everitt, "A Handbook of Statistical Analyses using S-PLUS," Chapman & Hall, 1994.
- [2] Paul Newbold, "Statistics for Business dan Economics," Printice Hall New Jersey, 1990.
- [3] Wijaya,IR, "Analisis Statistik dengan Program SPSS 10.0," Alfa Beta Bandung, 2001.