Tugas 3 Regresi Linear Berganda dan Seleksi Variabel

Nama : Rosa Amalia Nursinta (11190940000041)
Kelas : Matematika 5B

Analisis regresi linear berganda dan seleksi variable menggunakan uji-F untuk mendapatkan model terbaik

Data yang akan diolah

Data5

y	x1	x2	x3	x4	x5
1.95	1.31	1.07	0.44	0.75	0.35
2.9	1.55	1.49	0.53	0.9	0.47
0.72	0.99	0.84	0.34	0.57	0.32
0.81	0.99	0.83	0.34	0.54	0.27
1.09	1.05	0.9	0.36	0.64	0.3
1.22	1.09	0.93	0.42	0.61	0.31
1.02	1.08	0.9	0.4	0.51	0.31
1.93	1.27	1.08	0.44	0.77	0.34
0.64	0.99	0.85	0.36	0.56	0.29
2.08	1.34	1.13	0.45	0.77	0.37
1.98	1.3	1.1	0.45	0.76	0.38
1.9	1.33	1.1	0.48	0.77	0.38
8.56	1.86	1.47	0.6	1.01	0.65
4.49	1.58	1.34	0.52	0.95	0.5
8.49	1.97	1.59	0.67	1.2	0.59
6.17	1.8	1.56	0.66	1.02	0.59
7.54	1.75	1.58	0.63	1.09	0.59
6.36	1.72	1.43	0.64	1.02	0.63
7.63	1.68	1.57	0.72	0.96	0.68
7.78	1.75	1.59	0.68	1.08	0.62
10.15	2.19	1.86	0.75	1.24	0.72
6.88	1.73	1.67	0.64	1.14	0.55

Langkah analisis :

Pemodelan regresi linear berganda untuk mendapatkan model terbaik

1. Mendefinisikan matriks Data5 : memasukkan tiap elemen x1 sampai x5 ke matriks X
2. Menduga parameter dengan pendekatan matriks : membentuk model regresi berdasarkan nilai koefisien parameter yang diperoleh.
3. Uji-F dan Uji-t dengan pendekatan matriks : Uji-F digunakan untuk melihat apakah variabel independent secara versama-sama mempengaruhi variabel dependen. Uji-t digunakan untuk melihat apakah tiap-tiap variabel independent berpengaruh secara signifikan terhadap

variabel y atau tidak. Model yang baik adalah model yang tiap variabel bebas mempengaruhi variabel y.

4. Hitung R-squared dan Rsquared adjusted dengan pendekatan matriks : nilai R-squared dapat memperlihatkan seberapa besar pengaruh variabel x terhadapt varibel y.
5. Seleksi variabel menggunakan Uji-F : menseleksi variabel dengan metode menghilangkan satu persatu variabel bebas untuk mendapatkan model terbaik dengan memperhatikan nilai R-squared dan AIC yang diperoleh.
6. Kesimpulanan dan menentukan model terbaik : membentuk model terbaru yang sudah terseleksi dan juga melihat R-squared nya untuk mengetahui seberapa baik model ini.

Tabel Anova Data5

```
> #Regresi
> Data5 <- read.csv("D:/Data5.csv", sep=",")
> head(Data5)
     y   x1   x2   x3   x4   x5
1 1.95 1.31 1.07 0.44 0.75 0.35
2 2.90 1.55 1.49 0.53 0.90 0.47
3 0.72 0.99 0.84 0.34 0.57 0.32
4 0.81 0.99 0.83 0.34 0.54 0.27
5 1.09 1.05 0.90 0.36 0.64 0.30
6 1.22 1.09 0.93 0.42 0.61 0.31
> model5=lm(formula=y~., data=Data5)
> summary(model5)

Call:
lm(formula = y ~ ., data = Data5)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2610 -0.5373  0.1355  0.5120  0.8611

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.5122     0.9336  -6.976 3.13e-06 ***
x1            1.9994     2.5733   0.777  0.44851
x2           -3.6751     2.7737  -1.325  0.20378
x3            2.5245     6.3475   0.398  0.69610
x4            5.1581     3.6603   1.409  0.17791
x5           14.4012     4.8560   2.966  0.00911 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7035 on 16 degrees of freedom
Multiple R-squared:  0.9633,   Adjusted R-squared:  0.9519
F-statistic: 84.07 on 5 and 16 DF,  p-value: 6.575e-11

> anova(model5)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq  F value    Pr(>F)
x1         1 199.145 199.145 402.4397 9.131e-13 ***
x2         1   0.127   0.127   0.2560  0.619804
x3         1   4.120   4.120   8.3249  0.010765 *
x4         1   0.263   0.263   0.5325  0.476114
x5         1   4.352   4.352   8.7951  0.009109 **
Residuals 16   7.918   0.495
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

1. Matriks Data5

```
> y<-Data5$y
> X5<-as.matrix(cbind(x0=rep(1,nrow(Data5)), Data5[,-1])) #elemen
pada tiap x dimasukkan ke dalam matriks X5
> X5
      x0   x1   x2   x3   x4   x5
 [1,]  1 1.31 1.07 0.44 0.75 0.35
 [2,]  1 1.55 1.49 0.53 0.90 0.47
 [3,]  1 0.99 0.84 0.34 0.57 0.32
 [4,]  1 0.99 0.83 0.34 0.54 0.27
 [5,]  1 1.05 0.90 0.36 0.64 0.30
 [6,]  1 1.09 0.93 0.42 0.61 0.31
 [7,]  1 1.08 0.90 0.40 0.51 0.31
 [8,]  1 1.27 1.08 0.44 0.77 0.34
 [9,]  1 0.99 0.85 0.36 0.56 0.29
[10,]  1 1.34 1.13 0.45 0.77 0.37
[11,]  1 1.30 1.10 0.45 0.76 0.38
[12,]  1 1.33 1.10 0.48 0.77 0.38
[13,]  1 1.86 1.47 0.60 1.01 0.65
[14,]  1 1.58 1.34 0.52 0.95 0.50
[15,]  1 1.97 1.59 0.67 1.20 0.59
[16,]  1 1.80 1.56 0.66 1.02 0.59
[17,]  1 1.75 1.58 0.63 1.09 0.59
[18,]  1 1.72 1.43 0.64 1.02 0.63
[19,]  1 1.68 1.57 0.72 0.96 0.68
[20,]  1 1.75 1.59 0.68 1.08 0.62
[21,]  1 2.19 1.86 0.75 1.24 0.72
[22,]  1 1.73 1.67 0.64 1.14 0.55
>
>
```

2. Pendugaan Parameter mengggunakan pendekatan matriks

```
> #1. Duga koefisien parameter
> beta_hat<-(solve(t(X5) %*% X5)) %*% (t(X5) %*% y) #Tuliskan model reg
resi dan interpretasikan
> beta_hat
         [,1]
x0 -6.512215
x1  1.999413
x2 -3.675096
x3  2.524486
x4  5.158082
x5 14.401162
>
>
```

Pembahasan dan interpretasi :

1. Model Regresi
   - Berdasarkan output yang dihasilkan didapat nilai koefisien b0, b1, b2, b3, b4, b5 berturut-turut adalah -6.512215, 1.999413, -3.675096, 2.524486, 5.158082, 14.401162
   - Persamaan regresi yang diperoleh

     $\hat{Y} = -6.512215 + 1.999413X_1 - 3.675096X_2 + 2.524486X_3 + 5.158082X_4 + 14.401162X_5$
   - Hal ini berarti :
     - $b_1 = 1.999413$ menunjukkan bahwa ketika terjadi perubahan satu satuan X_1 maka akan meningkatkan rata-rata Y sebesar 1.999413, dengan asumsi X_2, X_3, X_4, X_5 tetap.
     - $b_2 = -3.675096$ menunjukkan bahwa ketika terjadi perubahan satu satuan X_2 maka akan menurunkan rata-rata Y sebesar 3.675096 ,dengan asumsi X_1, X_3, X_4, X_5 tetap.
     - $b_3 = 2.524486$ menunjukkan bahwa ketika terjadi perubahan satu satuan X_3 maka akan meningkatkan rata-rata Y sebesar 2.524486, dengan asumsi X_1, X_2, X_4, X_5 tetap.

- $b_4 = 5.158082$ menunjukkan bahwa ketika terjadi perubahan satu satuan X_4 maka akan meningkatkan rata-rata Y sebesar 5.158082, dengan asumsi X_1, X_2, X_3, X_5 tetap.
- $b_5 = 14.401162$ menunjukkan bahwa ketika terjadi perubahan satu satuan X_5 maka akan meningkatkan rata-rata Y sebesar 14.401162, dengan asumsi X_1, X_2, X_3, X_4 tetap.

3. Uji-F dengan pendekatan matriks

```
> #Uji-F dan Uji-t
> # ANOVA Table >> SSregfullmodel (R(beta)) = y'Xb; SSreg = y'Xb - (Totalyi)^2/
n; SSres = y'y - y'Xb
> SSreg<-(t(y)%*% (X5 %*% beta_hat)) - ((sum(y)^2)/length(y))
> dbreg<-ncol(X5)-1
> SSres<-(t(y) %*% (y)) - (t(y)%*% (X5 %*% beta_hat))
> dbres<-nrow(X5)-ncol(X5)
> MSreg<-SSreg/dbreg
> MSres<-SSres/dbres
> Fhit_mod5<-MSreg/MSres
> SSreg
         [,1]
[1,] 208.0072
> dbreg
[1] 5
> SSres
         [,1]
[1,] 7.917523
> dbres
[1] 16
> MSreg
         [,1]
[1,] 41.60145
> MSres
          [,1]
[1,] 0.4948452
> Fhit_mod5
         [,1]
[1,] 84.06962
>
>
> #Ftable
> Ftable_mod5<-qf(0.05, dbreg, dbres, lower.tail = FALSE, log.p = FALSE)
> Ftable_mod5
[1] 2.852409
>
> # keputusan berdasarkan Uji-F
> ifelse(Fhit_mod5 >= Ftable_mod5, "model layak", "model tidak layak")# interpre
tasikan
     [,1]
[1,] "model layak"
>
>
```

Pembahasan dan Interpretasi :

2. Uji-F
    - Hipotesis uji :

      H0 : $\beta = 0$ (model tidak layak)

      H1 : $\beta \neq 0$ (model layak)
    - α : $5\% = 0.05$
    - Statistik uji :

      $SS_{reg} = 208.0072$

      $SS_{res} = 7.917523$

      $MS_{reg} = 41.60145$

      $MS_{res} = 0.4948452$

      $F_{5,16,\alpha=0.05} = 2.852409$

      $F_{\text{hitung}} = 84.06962$
    - Daerah kritis :

      F_{tabel} yang diperoleh untuk $F_{5,16,\alpha=0.05}$ adalah 2.852409

      H_0 ditolak jika $F_{\text{hitung}} > F_{\text{tabel}}$

      Keputusan :

      F_{hitung} : 84.07, $F_{5,16,\alpha=0.05}$: 2.852409

      $F_{\text{hitung}} > F_{\text{tabel}}$, maka H_0 ditolak
    - Kesimpulan :

      Berdasarkan Uji-F memperlihatkan bahwa cukup bukti untuk variable X_1, X_2, X_3, X_4, X_5 secara bersama-sama berpengaruh terhadap variable Y

4. Uji t dengan pendekatan matriks

```
> # Uji-t
> s=sqrt(MSres)
> c_ji<-solve(t(X5) %*% X5)
>
> sbj<-function(c){
+   sbj<-sqrt(c)*s
+   sbj
+ }
>
> sb1<-sbj(c_ji[2,2])
> sb2<-sbj(c_ji[3,3])
> sb3<-sbj(c_ji[4,4])
> sb4<-sbj(c_ji[5,5])
> sb5<-sbj(c_ji[6,6])
> sb<-c(sb1,sb2,sb3,sb4,sb5)
> sb
[1] 2.573338 2.773660 6.347495 3.660283 4.855994
>
> # hipotesis beta=0
> thit<-function(b,stdb){
+   t<-b/stdb
+   t
+ }
>
> tb1<-thit(beta_hat[2],sb[1])
> tb2<-thit(beta_hat[3],sb[2])
> tb3<-thit(beta_hat[4],sb[3])
> tb4<-thit(beta_hat[5],sb[4])
> tb5<-thit(beta_hat[6],sb[5])
> t_stat<-c(tb1,tb2,tb3,tb4,tb5)
> t_stat
[1]  0.7769726 -1.3249989  0.3977137  1.4092029  2.9656468
>
> # ttable
> t.table<-qt(0.05,dbres, lower.tail = FALSE)
> t.table
[1] 1.745884
>
> # keputusan berdasarkan ujji-t
> ifelse(t_stat > t.table, "signifikan", "tidak signifikan") # interpretasi
kan
[1] "tidak signifikan" "tidak signifikan" "tidak signifikan" "tidak signifi
kan" "signifikan"
>
>
```

Pembahasan dan Interpretasi :

3. Uji-t
    - Hipotesis uji :

      $H_0 : \beta = 0$ (model tidak berpengaruh signifikan)

      $H_1 : \beta \neq 0$ (model berpengaruh signifikan)
    - $\alpha : 5\% = 0.05$ dan $\alpha : 1\% = 0.01$ (untuk x_5)
    - Statistik uji :

      $Sb_1 = 2.573338$, $Sb_2 = 2.773660$, $Sb_3 = 6.347495$, $Sb_4 = 3.660283$, $Sb_5 = 4.855994$

      $t_{hit1} = 0.7769726$, $t_{hit2} = -1.3249989$, $t_{hit3} = 0.3977137$, $t_{hit4} = 1.4092029$, $t_{hit5} = 2.9656468$

      $t_{16,\alpha=0.05} = 1.74588$, $t_{16,0.005} = 2.583487$
    - Daerah kritis :

      t_{tabel} yang diperoleh untuk $t_{16,0.025}$ adalah 1.745884 dan $t_{16,0.005}$ adalah 2.583487

      H_0 ditolak jika $t_{hitung} > t_{tabel}$
    - Keputusan :

      $t_{16,0.025} : 1.745884$, $t_{16,0.005} = 2.583487$

      $t_{hit1} = 0.7769726$, $t_{hitung} < t_{tabel}$ di $\alpha : 5\% = 0.05$, maka H_0 diterima

      $t_{hit2} = -1.3249989$, $t_{hitung} < t_{tabel}$ di $\alpha : 5\% = 0.05$, maka H_0 diterima

      $t_{hit3} = 0.3977137$, $t_{hitung} < t_{tabel}$ di $\alpha : 5\% = 0.05$, maka H_0 diterima

      $t_{hit4} = 1.4092029$, $t_{hitung} < t_{tabel}$ di $\alpha : 5\% = 0.05$, maka H_0 diterima

      $t_{hit5} = 2.9656468$, $t_{hitung} > t_{tabel}$ di $\alpha : 1\% = 0.01$, maka H_0 ditolak

    - Kesimpulan :
    - Berdasarkan Uji-t menunjukkan bahwa dari serangkaian X_1, X_2, X_3, X_4, sampai X_5 yang berpengaruh signifikan terhadap variable Y hanyalah X5 pada $\alpha : 5\% = 0.05$ maupun $\alpha : 1\% = 0.01$.

5. R-squared dan R-squared adjusted dengan pendekatan matriks

```
> # R-squared
> SStot<-SSreg+SSres
> R_sq<-SSreg/SStot*100
> R_sq # interpretasikan : sebesar 96% model mampu menjelaskan keragaman va
riabel y, dan sekitar 3.7% keragaman yang tidak mampu dijelaskan oleh varia
bel penjelas dan terletak pada komponen error
         [,1]
[1,] 96.3332
>
>
> # R-squared adjusted
> n<-nrow(Data5)
> k<-ncol(X5) - 1
> R_sq.adj <- 1-(((n-1)/(n-k-1))*(SSres/SStot))
> R_sq.adj #interpretasikan
          [,1]
[1,] 0.9518733
>
>
```

Pembahasan dan interpretasi :

4. R-squared san R-aquared adjusted
   - $SSreg = 208.0072$

     $SSres = 7.917523$

     $SStot = 215.9248$
   - R-squared = 96.3332
   - R-squared adjusted = 95.18733
   - Kesimpulan :

     Sebesar 96% model mampu menjelaskan keragaman variabel y dan sekitar 3,7% keragaman yang tidak mampu dijelaskan oleh variabel penjelas dan terletak pada komponen error.

6. Seleksi variable menggunakan Uji-F

```
> # Seleksi Variabel
>
> #Uji-F Parsial (Backward elimination) menghilangkan satu per satu variabel p
enjelas
> full.model <- lm(y~., data=Data_Modlin)
> reduced.model <- step(full.model, direction="backward")
Start:  AIC=-10.48
y ~ x1 + x2 + x3 + x4 + x5

       Df Sum of Sq     RSS      AIC
- x3    1    0.0783  7.9958 -12.2668
- x1    1    0.2987  8.2163 -11.6684
<none>               7.9175 -10.4832
- x2    1    0.8688  8.7863 -10.1927
- x4    1    0.9827  8.9002  -9.9093
- x5    1    4.3522 12.2697  -2.8460

Step:  AIC=-12.27
y ~ x1 + x2 + x4 + x5

       Df Sum of Sq     RSS     AIC
- x1    1    0.2856  8.2814 -13.495
<none>               7.9958 -12.267
- x2    1    0.8193  8.8151 -12.121
- x4    1    0.9869  8.9827 -11.706
- x5    1    8.6436 16.6394   1.856

Step:  AIC=-13.49
y ~ x2 + x4 + x5

       Df Sum of Sq     RSS      AIC
- x2    1    0.6978  8.9793 -13.7148
<none>               8.2814 -13.4946
- x4    1    2.8116 11.0931  -9.0639
- x5    1   14.1791 22.4606   6.4558

Step:  AIC=-13.71
y ~ x4 + x5

       Df Sum of Sq     RSS      AIC
<none>               8.9793 -13.7148
- x4    1    2.7879 11.7671  -9.7661
- x5    1   15.6948 24.6740   6.5236
> reduced.model$anova #interpretasikan
  Step Df   Deviance Resid. Df Resid. Dev       AIC
1      NA         NA        16   7.917523 -10.48321
2 - x3  1 0.07827274        17   7.995795 -12.26679
3 - x1  1 0.28564666        18   8.281442 -13.49456
4 - x2  1 0.69780985        19   8.979252 -13.71477
>
>
```

Pembahasan dan interpretasi :

5. Seleksi Variabel
    - Indikator kebaikan model :
        - Adjusted R^2 : Semakin besar nilainya maka semakin baik modelnya
        - Nilai AIC (Akaike Information Criterion) : Semakin kecil nilainya maka semakin baik modelnya
    - Berdasarkan output yang dihasilkan diperoleh nilai AIC tiap seleksi variable yaitu
        - AIC = -10.48 untuk $y = x1 + x2 + x3 + x4 + x5$
        - AIC = -12.27 untuk $y = x1 + x2 + x4 + x5$
        - AIC = -13.49 untuk $y = x2 + x4 + x5$
        - AIC = -13.71 untuk $y = x4 + x5$
    - Kesimpulan :

      Berdasarkan nilai AIC yang diperoleh dapat disimpulkan bahwa model terbaik untuk Data5 adalah $y = x4 + x5$. Dan variable yang tereliminasi yaitu x1, x2, dan x3.

7. Kesimpulan dan menentukan model terbaik

```
> # Model terbaik
> model_final=lm(formula=y~x4+x5, data=Data5)
> summary(model_final)

Call:
lm(formula = y ~ x4 + x5, data = Data5)

Residuals:
     Min       1Q   Median       3Q      Max
-1.56123 -0.49604  0.09069  0.45717  0.98057

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.3351     0.6009 -10.543 2.23e-09 ***
x4            4.1542     1.7104   2.429   0.0252 *
x5           15.0160     2.6057   5.763 1.49e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6875 on 19 degrees of freedom
Multiple R-squared:  0.9584,   Adjusted R-squared:  0.954
F-statistic: 218.9 on 2 and 19 DF,  p-value: 7.584e-14

> anova(model_final)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x4         1 191.251 191.251  404.68 2.865e-14 ***
x5         1  15.695  15.695   33.21 1.491e-05 ***
Residuals 19   8.979   0.473
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> ## jadi x1, x2, x3 tidak berpengaruh signifikan terhadap y. maka dapat kit
a seleksi variabel-variabel tersebut dan didapatkan variabel terbaik yaitu x
4, x5
>
```

Kesimpulan :

- Setalah melakukan seleksi variable, didapat model regresi yang terbaik yaitu $\hat{Y} = -6.3351 + 4.1542X4 + 15.0160X5$.
- Hal ini berarti :
    - $b_4 = 4.1542$ menunjukkan bahwa ketika terjadi perubahan satu satuan X_4 maka akan meningkatkan rata-rata Y sebesar 4.1542, dengan asumsi X_5 tetap.

- $b_5 = 15.0160$ menunjukkan bahwa ketika terjadi perubahan satu satuan X_5 maka akan meninngkatkan rata-rata Y sebesar 15.0160, dengan asumsi X_4 tetap.

- Dengan R-squared dan R-squared adjusted yang baru berturut-turut yaitu 0.9584 dan 0.954, artinya sebesar 95% model ini mampu menjelaskan keragaman variable y, dan sekitar 4% keragaman yang tidak mempu dijelaskan oleh variable x4 dan x5 dan terletak pada komponen error.