PENGGUNAAN KRITERIA rc_P PADA PEMILIHAN PEUBAH BEBAS TERBAIK JIKA TERDAPAT MULTIKOLINEARITAS

Harmi Sugiarti (harmi@mail.ut.ac.id)
Universitas Terbuka

ABSTRACT

Some procedures can be used for selecting independent variables, one of them is the procedure of all possible regression with robust C_p (RC_p) criterion. This statistic is not sensitive with multicollinearity in model and outlier residuals. The aim of this article is to investigate the use of RC_p criterion in selecting independent variables. The result of the simulation experimental data shows that the RC_p criterion fits enough to select independent variables.

Keywords: independent variables selection, robust, multicolinearity

Masalah pemilihan peubah bebas yang sesuai sering merupakan masalah yang kompleks, hal ini disebabkan karena peubah-peubah bebas yang kita miliki belum tentu merupakan peubah-peubah bebas yang diperlukan untuk pemodelan. Tidak adanya prosedur matematis yang dapat menghasilkan sebuah penyelesaian yang baik untuk permasalahan pemilihan peubah, mengakibatkan prosedur yang tersedia hanya merupakan metode untuk menerangkan struktur data yang ada atau eksplorasi (Montgomery & Peck, 1992).

Salah satu prosedur yang dapat digunakan adalah prosedur semua kemungkinan regresi yang memuat peubah bebas potensial dan memilih persamaan terbaik dengan kriteria C_p-*Mallows* (Draper & Smith, 1981). Prosedur ini dilakukan pada kondisi ideal, yakni antara lain dipenuhinya asumsi tidak adanya korelasi diantara peubah-peubah bebas (multikolinearitas) dalam model regresi linear berganda.

Asumsi tidak adanya multikolinearitas diperlukan oleh metode kuadrat terkecil *(Ordinary Least Square, OLS)* untuk mendapatkan penaksir parameter yang bersifat tak bias linear terbaik *(Best Linear Unbiased Estimator, BLUE)* dari model regresi linear berganda $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi} + \varepsilon_i, \quad i = 1,2,...,n$ dengan Y_i adalah nilai peubah respons pada pengamatan ke-i, X_i adalah nilai peubah bebas pada pengamatan ke-i dan $\beta_0, \beta_1, ..., \beta_p$ adalah koefisien regresi yang tidak diketahui nilainya.

Meskipun nilai penduga $\hat{\sigma}^2$ kecil, adanya multikolinearitas mengakibatkan masalah pendugaan parameter regresi, yakni besarnya nilai variansi penduga $\hat{\beta}_j$ $\left(\text{var}(\hat{\beta}_j)\right)$, sehingga keadaan ini akan menyebabkan uji statistik secara parsial untuk koefisien garis regresi tidak signifikan.

Indikasi adanya masalah kolinearitas ditunjukkan dengan suatu diagnostik terhadap besarnya nilai *variance inflation factor (VIF)* dari $\hat{\beta}_j$. *VIF* adalah suatu faktor yang mengukur

seberapa besar kenaikan variansi dari penduga $\hat{\beta}_j$ dibandingkan terhadap peubah bebas lain yang saling orthogonal atau bebas. Misalkan R_j^2 adalah nilai koefisien determinasi dari peubah bebas X_j jika diregresikan terhadap semua peubah bebas X yang lainnya, maka nilai *VIF* dinyatakan sebagai $VIF = \frac{1}{1-R_j^2}$. Nilai *VIF* yang besar $(VIF > 10)$ bisa digunakan sebagai petunjuk adanya multikolinearitas (Neter & Wasserman, 1990).

Penanganan kasus multikolinearitas dalam model, kadangkala diikuti munculnya penyimpangan asumsi lainnya, diantaranya munculnya pengamatan pencilan *(outlier)* dalam data. Adanya pengamatan pencilan dalam data, dapat mengakibatkan penaksir koefisien garis regresi yang diperoleh tidak tepat. Namun demikian tindakan membuang (menolak) begitu saja suatu pengamatan pencilan bukanlah tindakan yang bijaksana, karena adakalanya pengamatan pencilan memberikan informasi yang cukup berarti.

Dalam hal tidak dipenuhinya asumsi tidak ada multikolinearitas, salah satu metode alternatif yang dapat dicoba adalah prosedur semua kemungkinan regresi yang memuat peubah bebas potensial, dan memilih persamaan terbaik dengan menggunakan kriteria $RC_p\left(robust\ C_p\right)$ (Sommer & Huggins, 1996).

Menurut Staudte & Sheather (1990), jika hubungan linear antara satu peubah respon dengan peubah-peubah bebasnya dimodelkan sebagai $Y_i = X_i^T\beta + \varepsilon_i$, dimana X_i^T menyatakan baris ke-i dari matriks rancangan *X*, β menyatakan parameter model dan ε_i menyatakan suku galat. Penduga kemungkinan maksimum *(M-estimator)* $\hat{\beta}_p$ untuk model dengan *p* parameter diperoleh dengan cara meminimumkan $\sum_i \rho\left(x_i, e_i\right) = \sum_i \rho\left(x_i, y_i - x_i^T\hat{\beta}_\rho\right)$ atau mencari penyelesaian dari persamaan $\sum_i x_i\ \eta\left(x_i, y_i - x_i^T\hat{\beta}_\rho\right) = 0$, dengan $\eta\left(x,e\right) = \rho'\left(x,e\right)$ untuk berbagai fungsi konveks $\rho\left(x,e\right)$ yang dapat diturunkan dan memenuhi $\eta\left(x,0\right) = 0$. Karena penduga $\hat{\beta}_p$ yang diperoleh ini bukan merupakan skala *invariant*, yaitu jika sisaannya $\left(e_i = y_i - x_i^T\hat{\beta}\right)$ digandakan dengan suatu konstanta akan diperoleh penyelesaian yang tidak sama seperti sebelumnya. Sehingga untuk mendapatkan skala *invariant*, digunakan nilai $\frac{e_i}{\sigma}$ sebagai pengganti e_i, dengan σ adalah faktor skala yang juga perlu diduga. Dengan demikian persamaan yang ada menjadi:

$$\sum_i x_i\ \Psi\left(x_i, \frac{e_i}{\sigma}\right) = \sum_i x_i\ \Psi\left(x_i, \frac{y_i - x_i^T\hat{\beta}_\eta}{\sigma}\right) = \sum_i x_i\left(y_i - x_i^T\hat{\beta}_\eta\right)w_i = 0$$

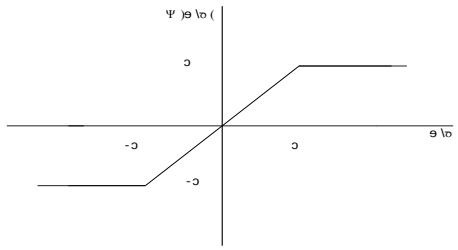
dengan fungsi pembobot $w_i = w\left(x_i, \frac{y_i - x_i^T\hat{\beta}_\eta}{\sigma}\right) = \frac{\Psi\left(x_i, \frac{e_i}{\sigma}\right)}{\frac{e_i}{\sigma}}$ yang bernilai antara 0 dan 1. Secara umum fungsi pembobot ditulis sebagai berikut.

$$w_i = w\left(x_i, \frac{y_i - x_i^T\hat{\beta}_\eta}{\sigma}\right) = w\left(x_i, \frac{e_i}{\sigma}\right) = \frac{\sigma\nu(x_i)}{e_i}\psi_c\left(\frac{e_i}{\sigma\nu(x_i)}\right),$$

dengan Ψ_c adalah *influence function* dan $v(x_i)$ adalah suatu fungsi yang tergantung pada *x* melalui *leverage*-nya. Dalam hal ini ditentukan nilai $v(x_i) = \frac{(1-h_{ii})}{\sqrt{h_{ii}}}$ dan $\hat{\sigma} = s_{(i)}$ serta fungsi Huber dengan bentuk:

$$\Psi_c(\frac{e}{\sigma}) = \begin{cases} c, & \text{jika } \frac{e}{\sigma} > c \\ \frac{e}{\sigma}, & \text{jika } \left|\frac{e}{\sigma}\right| \le c \\ -c, & \text{jika } \frac{e}{\sigma} < -c \end{cases}$$

Perhatikanlah grafik fungsi $\Psi_c(\frac{e}{\sigma})$ pada Gambar 1.

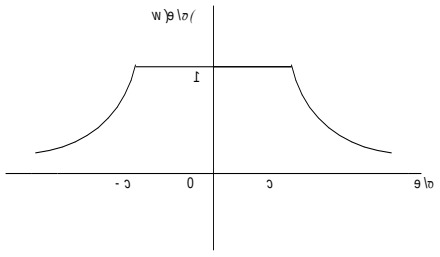


Gambar 1. Fungsi Huber

Nilai pembobot w_i menjadi tergantung pada kombinasi besarnya *leverage* dan *studentized residual* melalui *difference in the fitted value-standardized (DFFITS)*. Secara singkat nilai pembobot w_i dapat dinyatakan dalam bentuk:

$$w\left(x_i, \frac{y_i - x_i^T \hat{\beta}_\eta}{\sigma}\right) = w\left(x_i, \frac{e_i}{\sigma}\right) = \min\left(\frac{2\sqrt{p/n}}{|DFFITS_i|}, 1\right)$$

Perhatikanlah grafik fungsi $w\left(x_i, \frac{e_i}{\sigma}\right)$ pada Gambar 2.



Gambar 2 . Fungsi Pembobot Huber

Jadi, persamaan $\sum_i \left(y_i - x_i^T \hat{\beta}_\eta\right) w_i x_i = 0$ dapat dituliskan dalam bentuk matriks $X^T W X \beta = X^T W Y$ yang kita kenal sebagai persamaan normal kuadrat terkecil tertimbang dengan *W* adalah matriks diagonal yang berisi pembobot. Solusi persamaan normal tersebut akan memberikan dugaan untuk β yaitu $\hat{\beta} = \left(X^T W X\right)^{-1} X^T W Y$ dan penduga-M untuk β diperoleh dengan cara melakukan iterasi sampai diperoleh suatu hasil yang kovergen, cara ini biasa dikenal sebagai metode kuadrat terkecil tertimbang secara iteratif *(iteratively reweighted least square).*

Ronchetti & Staudte (1994) memberikan statistik RC_p sebagai kriteria dalam masalah pemilihan persamaan regresi terbaik berdasarkan pembobot $\hat{w}_i$ dan penduga-M parameter $\hat{\beta}_p$. Statistik RC_p untuk persamaan regresi *P* adalah:

$$RC_p = \frac{W_p}{\hat{\sigma}^2} - (U_p - V_p)$$

dengan

$$W_p = \sum_i \hat{w}_i^2 e_i^2 = \sum_i \hat{w}_i^2 \left(y_i - x_i^T \hat{\beta}_p\right)^2$$

$$U_p = \sum_i \text{var}\left(\hat{w}_i e_i\right) = \sum_i \text{var}\left[\hat{w}_i \left(y_i - x_i^T \hat{\beta}_p\right)\right]$$

$$V_p = \sum_i \text{var}\left[\hat{w}_i x^T \left(\hat{\beta}_p - \beta\right)\right]$$

$$\hat{\sigma}^2 = \frac{W_{full}}{U_{full}}$$

U_p & V_p dihitung dengan asumsi bahwa submodel *P* benar dan $\sigma = 1$. Untuk memilih suatu persamaan regresi terbaik dapat dilakukan dengan melihat plot antara RC_p dan V_p. Model dengan dengan nilai RC_p dibawah persamaan garis $RC_p = V_p$ dapat dipilih sebagai model terbaik. Jika penduga yang diperoleh merupakan penduga bentuk Huber, nilai V_p akan mendekati *p*.

Tulisan ini bertujuan untuk mengkaji penggunaan kriteria RC_p dalam menentukan peubah bebas *terbaik* jika terdapat multikolinearitas dalam model regresi linear berganda.

METODOLOGI

Data yang dipergunakan dalam penelitian ini terdiri dari *data simulasi*, yaitu data yang dibangkitkan dengan bantuan program MINITAB. Sebanyak 40 pasang data yang dibangkitkan adalah data peubah bebas $\left(X_1, X_2, X_3, X_4\right)$, data galat $\left(\varepsilon\right)$ dan data peubah tak bebas $\left(Y\right)$ yang diperoleh melalui asumsi model $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$, $\beta_0 = 5$, $\beta_1 = 1$, $\beta_2 = 1$. Guna menunjang pemahaman, digunakan *data eksperimen*, yaitu data kadar *tar* $\left(X_1\right)$, *nicotine* $\left(X_2\right)$, *carbon monoxide* $\left(Y\right)$ dalam rokok, dan berat rokok $\left(X_3\right)$. Kadar *tar*, *nicotine*, dan *carbon monoxide* diukur dalam mg, dan berat rokok dalam g (McClave & Sincich, 2003).

Berdasarkan data yang ada, ditentukan peubah tak bebas terbaik dalam model regresi linear berganda dengan kriteria C_p-*Mallows* dan RC_p serta, selanjutnya membandingkan hasil yang diperoleh.

HASIL DAN PEMBAHASAN

Hasil simulasi memberikan empat puluh pasang data (Y, X_1, X_2, X_3, X_4) dengan hasil analisis variansi dan analisis regresi dapat dilihat pada Tabel 1 berikut ini.

Tabel 1. Analisis Regresi dan Analisis Variansi untuk Data Simulasi

Regression Analysis

The regression equation is

y = 5.32 + 0.996 x1 + 1.05 x2 - 0.0521 x3 - 0.00456 x4

Predictor	Coef	StDev	T	P	VIF
Constant	5.3156	0.6015	8.84	0.000	
x1	0.995980	0.006687	148.93	0.000	1.1
x2	1.05404	0.03376	31.22	0.000	28.7
x3	-0.05213	0.03347	-1.56	0.128	28.3
x4	-0.004563	0.006513	-0.70	0.488	1.0

S = 1.177 R-Sq = 99.9% R-Sq(adj) = 99.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	53002	13250	9559.28	0.000
Error	35	49	1		
Total	39	53050			

Source	DF	Seq SS
x1	1	19730
x2	1	33268
x3	1	3
x4	1	1

Durbin-Watson statistic = 1.73

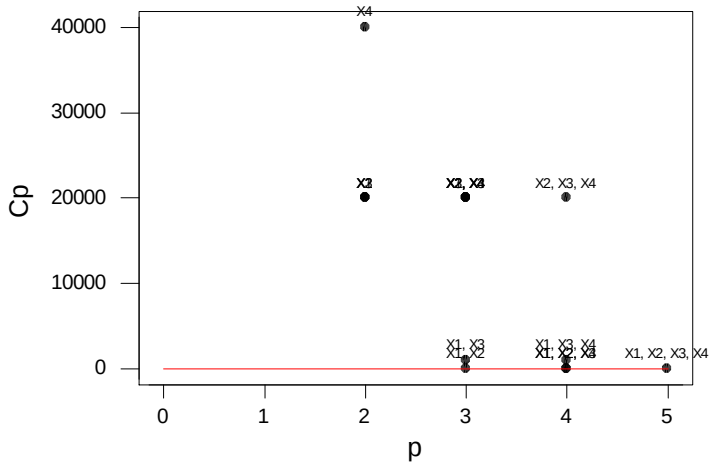
Selain dari besarnya nilai koefisien determinasi $(R^2 = 0,999)$, secara simultan uji F memberikan nilai $(F = 9559,28)$ yang sangat signifikan dengan $(p - value = 0,00)$, sehingga dapat dikatakan bahwa model regresi linear sudah sesuai untuk data yang ada. Secara parsial, uji t memberikan nilai yang tidak signifikan $(t = -1,56 \; ; \; p - value = 0,128)$ untuk peubah bebas X_3 dan $(t = -0,70 \; ; \; p - value = 0,488)$ untuk peubah bebas X_4. Hal ini kemungkinan disebabkan karena peubah X_2 berkorelasi positif dengan peubah bebas X_3, yakni dengan koefisien korelasi sebesar $r_{23} = 0,982$. Adanya multikolinearitas juga dapat dilihat dari nilai faktor pembesar variansi $(VIF > 10)$, dimana peubah bebas X_2 mempunyai nilai $VIF = 28,7$ dan peubah bebas X_3 mempunyai nilai $VIF = 28,3$.

Nilai statistik C_p-*Mallows* dan RC_p untuk berbagai kombinasi peubah bebas dapat dilihat pada Tabel 2. Statistik C_p-*Mallows* memberikan rekomenda-si bahwa persamaan regresi dengan peubah bebas (X_1, X_2) , (X_1, X_2, X_3), atau (X_1, X_2, X_4) merupakan persamaan regresi terbaik, meskipun persamaan regresi dengan peubah bebas (X_1, X_2) atau (X_1, X_2, X_4) mempunyai bias sedikit lebih besar. Hal ini dapat dilihat dari nilai C_p-*Mallows* yang mendekati nilai *p*, namun nilai C_p-*Mallows* berada di atas garis $C_p = p$ (Gambar 3).

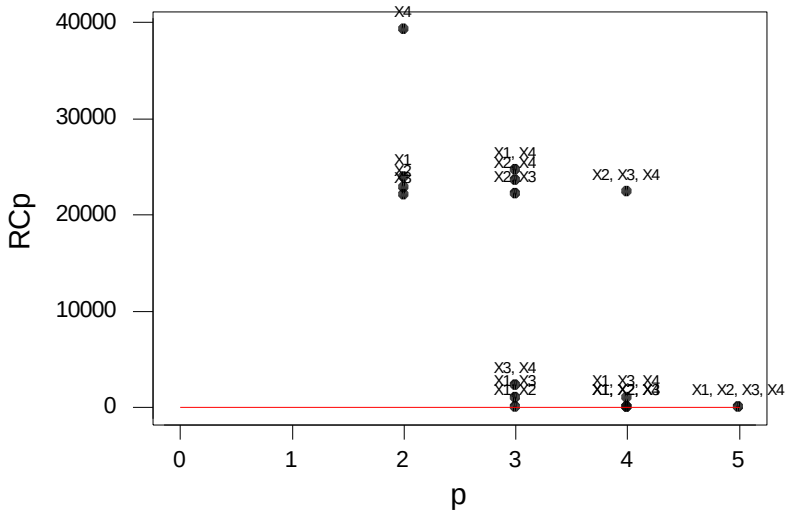
Tabel 2. Statistik C_p-*Mallows* dan RC_p untuk Data Simulasi

Variabel	p	C_p-*Mallows*	RC_p
X_1	2	2E+04	23923,7785
X_2	2	2E+04	22776,5362
X_3	2	2E+04	22063,2951
X_4	2	4E+04	39276,5874
X_1, X_2	3	3,8	4,5585
X_1, X_3	3	977,8	962,3766
X_1, X_4	3	2E+04	24701,4574
X_2, X_3	3	2E+04	22188,6034
X_2, X_4	3	2E+04	23585,6904
X_3, X_4	3	2E+04	22887,2803
X_1, X_2, X_3	4	3,5	3,9893
X_1, X_2, X_4	4	5,4	4,0350
X_1, X_3, X_4	4	977,7	941,4830
X_2, X_3, X_4	4	2E+04	22444,6311
X_1, X_2, X_3, X_4	5	5,0	4,9174

Statistik RC_p memberikan hasil yang tidak terlalu berbeda dengan statistik C_p-*Mallows*, yakni selain memberikan rekomendasi persamaan regresi dengan peubah bebas (X_1, X_2) atau (X_1, X_2, X_3) sebagai persamaan regresi terbaik, statistik RC_p juga memberikan rekomendasi bahwa persamaan regresi dengan peubah bebas (X_1, X_2, X_4) merupakan persamaan regresi terbaik. Seperti halnya pada statistik C_p-*Mallows*, persamaan regresi dengan peubah bebas (X_1, X_2) mempunyai bias sedikit lebih besar. Hal ini dapat dilihat dari nilai RC_p yang meskipun mendekati nilai *p*, namun nilai RC_p berada di atas garis $RC_p = p$ (Gambar 4).



Gambar 3 . Plot Statistik C_p terhadap p untuk Data Simulasi



Gambar 4 . Plot Statistik RC_p terhadap p untuk Data Simulasi

Analisis lebih jauh menunjukkan bahwa untuk persamaan regresi dengan dua peubah bebas di dalamnya, statistik C_p-*Mallows* lebih baik dibanding statistik RC_p dalam memberikan rekomendasi tentang peubah bebas yang masuk dalam model.

Sedangkan pada model dengan menggunakan tiga peubah bebas, statistik RC_p memberikan rekomendasi yang lebih baik dibanding statistik C_p-*Mallows* . Hal ini dapat dilihat dari nilai C_p-*Mallows* yang lebih memilih peubah bebas (X_1, X_2, X_3) dibanding peubah bebas (X_1, X_2, X_4), di mana kehadiran peubah bebas X_2 dan X_3 dalam model secara bersama-sama tidak diharapkan karena terdapat korelasi yang cukup tinggi antara peubah bebas X_2 dan X_3.

Berdasarkan data eksperimen berukuran 25 tentang kadar *tar* (X_1), *nicotine* (X_2), *carbon monoxide* (Y) dalam rokok, dan berat rokok (X_3) diperoleh hasil analisis variansi dan analisis regresi seperti pada Tabel 3.

Karena nilai koefisien determinasi besar $(R^2 = 0,919)$ dan secara simultan uji F memberikan nilai yang cukup signifikan $(F = 78,98)$ dengan $(p - value = 0,00)$, maka dapat dikatakan bahwa model regresi linear sudah sesuai untuk data yang ada. Secara parsial, uji t memberikan nilai yang tidak signifikan $(t = -0,67 \; ; \; p - value = 0,507)$ untuk peubah bebas *nicotine* (X_2) dan $(t = -0,03; \; p - value = 0,974)$ untuk peubah bebas berat rokok (X_3).

Tabel 3. Analisis Regresi dan Analisis Variansi untuk Data Hasil Eksperimen

REGEGRESSION ANALYSIS

The regression equation is
CO = 3.20 + 0.963 Tar - 2.63 Nicotine - 0.13 Berat

Predictor	Coef	StDev	T	P	VIF
Constant	3.202	3.462	0.93	0.365	
Tar	0.9626	0.2422	3.97	0.001	21.6
Nicotine	-2.632	3.901	-0.67	0.507	21.9
Berat	-0.130	3.885	-0.03	0.974	1.3

S = 1.446 R-Sq = 91.9% R-Sq(adj) = 90.7%

ANALYSIS OF VARIANCE

Source	DF	SS	MS	F	P
Regression	3	495.26	165.09	78.98	0.000
Error	21	43.89	2.09		
Total	24	539.15			

Source	DF	Seq SS
Tar	1	494.28
Nicotine	1	0.97
Berat	1	0.00

Unusual Observations

Obs	Tar	CO	Fit	StDev Fit	Residual	St Resid
3	29.8	23.500	26.393	1.030	-2.893	-2.85RX

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.
Durbin-Watson statistic = 2.86

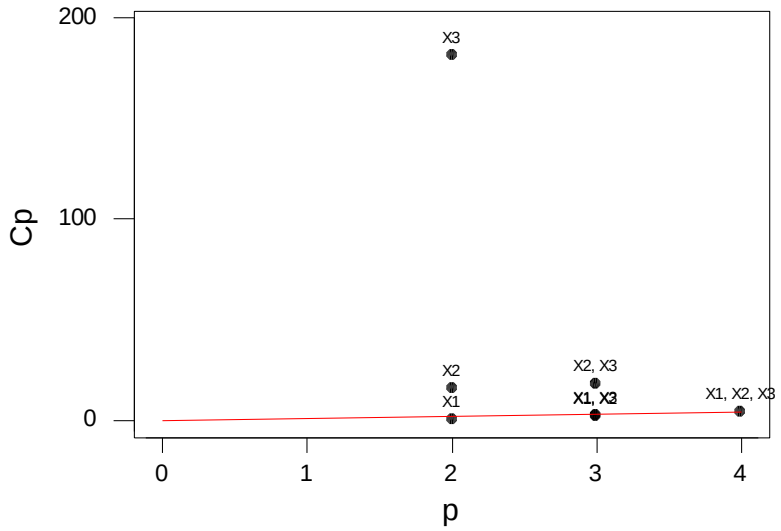
Jika dilihat besarnya koefisien korelasi antara peubah bebas *tar* (X_1) dan bebas *nicotine* (X_2) yakni $r_{12} = 0{,}977$, maka tidak signifikannya uji t secara parsial mungkin disebabkan karena adanya multikolinearitas, yang juga dapat dilihat dari nilai faktor pembesar variansi $(VIF > 10)$, dimana peubah bebas X_1 mempunyai nilai $VIF = 21{,}6$ dan peubah bebas X_2 mempunyai nilai $VIF = 21{,}9$.

Selain adanya multikolinearitas dalam model, Tabel 3 juga menunjukkan adanya gejala pencilan *(outlier)* yakni pengamatan ke-3. Nilai statistik C_p-*Mallows* dan RC_p untuk berbagai kombinasi peubah bebas dapat dilihat pada Tabel 4 sebagai berikut.

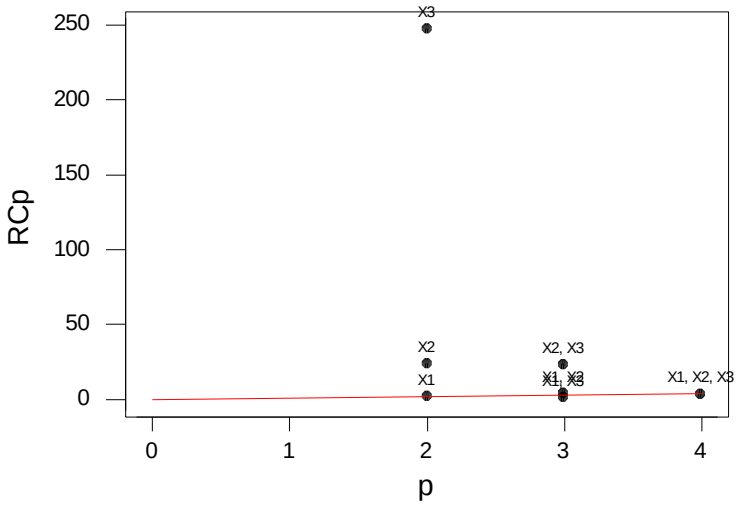
Tabel 4. Statistik C_p-*Mallows* dan RC_p untuk Data Hasil Eksperimen

Variabel	p	C_p-*Mallows*	RC_p
X_1	2	0,5	1,9322
X_2	2	15,8	24,2161
X_3	2	181,4	247,1458
X_1, X_2	3	2,0	4,0064
X_1, X_3	3	2,5	1,3482
X_2, X_3	3	17,8	23,3001
X_1, X_2, X_3	4	4,0	3,7432

Statistik C_p-*Mallows* memberikan rekomendasi bahwa persamaan regresi dengan peubah bebas (X_1), (X_1, X_2), atau (X_1, X_3) merupakan persamaan regresi terbaik. Hal ini dapat dilihat dari nilai C_p-*Mallows* yang mendekati nilai p (Gambar 5).



Gambar 5 . Plot Statistik C_p terhadap p untuk Data Hasil Eksperimen



Gambar 6 . Plot Statistik RC_p terhadap p untuk Data Hasil Eksperimen

Statistik RC_p memberikan hasil yang sedikit berbeda dengan statistik C_p-*Mallows*, yakni statistik RC_p hanya memberikan rekomendasi persamaan regresi dengan peubah bebas (X_1) atau (X_1, X_3) sebagai persamaan regresi terbaik, hal ini dapat dilihat dari nilai RC_p yang mendekati nilai *p* dan nilai RC_p berada di bawah garis $RC_p = p$ (Gambar 6).

Analisis lebih jauh menunjukkan bahwa untuk persamaan regresi dengan dua peubah bebas di dalamnya, statistik RC_p lebih baik dibanding statistik C_p-*Mallows* dalam memberikan rekomendasi tentang peubah bebas yang masuk dalam model. Hal ini dapat dilihat dari peubah bebas yang direkomendasikan, dimana statistik C_p-*Mallows* merekomendasikan persamaan regresi dengan peubah bebas (X_1, X_2) padahal diketahui bahwa koefisien korelasi peubah bebas *tar* (X_1) berkorelasi positif dengan peubah bebas *nicotine* (X_2), yakni dengan sebesar $r_{12} = 0,977$. Sehingga kehadiran peubah bebas X_1 dan X_2 dalam model secara bersama-sama tidak diharapkan.

KESIMPULAN

Secara keseluruhan dapat disimpulkan bahwa statistik RC_p memberikan rekomendasi yang tidak jauh berbeda dengan statistik C_p-*Mallows* tentang peubah bebas yang masuk dalam model, jika terdapat multikolinearitas dalam model regresi linear berganda. Statistik RC_p memberikan rekomendasi yang lebih baik dibandingkan dengan statistik C_p-*Mallows* tentang peubah bebas yang masuk dalam model jika terdapat multikolinearitas dan pencilan *(outlier)* dalam model regresi linear berganda.

REFERENSI

Draper,N.R. & Smith,H. (1981). *Applied regression analysis*. 2nd ed. New York: Wiley.

McClave, J.T. & Sincich,T. (2003). *Statistics*. 9th ed. New Jersey: Prentice-Hall.

Montgomery, D.C. & Peck, E.A. (1992). *Introduction to linear regression analysis*. 2nd ed. New York: Wiley.

Neter, J. & Wasserman,W. (1990). *Applied linear statistical models*. 3rd ed. Homewood, Illinois: Irwin.

Ronchetti, E. & Staudte, R.G. (1994). *A robust version of Mallow's C_p. J.Am. Statist.Ass*., 89, 550-559.

Sommer,S. & Huggins,R.M. (1996). *Variables selection using the wald test and a robust Cp*. *Appl. Statist.*, 45, 15-29.

Staudte,R.G. & Sheather,S.J. (1990). *Robust estimation and testing*. New York: Wiley.