RESUME STATISTIKA REGRESI

KELOMPOK 07

“VARIABLE SELECTION AND MODEL BUILDING”

DISUSUN OLEH :

1. Reza Putri Angga (22083010006)

2 .Kanessa Jasmine Prisheila A.S. (22083010016)

3 .Sharleen Agustine (22083010030)

DOSEN PENGAMPU :

Aviolla Terza Damaliana, S.Si., M.Stat (199408022022032015)

UNIVERSITAS PEMBANGUNAN NASIONAL “VETERAN”

JAWA TIMUR

2023

10.1 *INTRODUCTION* / PENDAHULUAN

10.1.1 *MODEL BUILDING PROBLEM* / MASALAH PEMBANGUNAN MODEL

Pada bab-bab sebelumnya, penekanan diberikan pada pentingnya variabel regresor yang diketahui dalam model, memastikan bentuk fungsional yang benar, dan menghindari pelanggaran asumsi dasar. Namun, terkadang pertimbangan teoritis atau pengalaman sebelumnya dapat membantu dalam memilih regresor. Pendekatan klasik untuk pemilihan model regresi digunakan dengan asumsi pemahaman yang baik tentang bentuk dasar model dan semua regresor yang relevan. Strateginya melibatkan penyesuaian model penuh, analisis komprehensif, penanganan kolinearitas, penilaian transformasi, penggunaan uji t, dan analisis residu menyeluruh untuk memastikan kecukupan model.

Masalah pemilihan variabel muncul dalam situasi praktis, terutama ketika melibatkan data historis, di mana ada banyak potensi regresor yang ada, tetapi hanya sedikit yang mungkin signifikan. Proses ini, sering disebut sebagai pemilihan variabel. Metode ini sangat penting terutama dalam keberadaan multikolinearitas. Meskipun pemilihan variabel adalah metode umum untuk mengatasi multikolinearitas, ini tidak menjamin penghilangannya. Terdapat beberapa kasus di mana regresor yang saling terkait secara erat seharusnya ada dalam model. Metode yang digunakan dalam pemilihan variabel membantu membenarkan inklusi regresor yang sangat terkait dalam model akhir.

Multikolinearitas bukan satu-satunya untuk alasan menggunakan teknik pemilihan variabel, bahkan hubungan yang ringan tidak terdeteksi oleh diagnosis multikolinearitas dapat memengaruhi pemilihan model. Teknik pemilihan model yang efektif meningkatkan kepercayaan pada model atau model akhir yang direkomendasikan. Membangun model regresi dengan subset regresor yang tersedia melibatkan tujuan yang saling bertentangan. Di satu sisi, kita ingin menyertakan sebanyak mungkin regresor untuk memengaruhi nilai yang diprediksi y, sementara di sisi lain, kita ingin meminimalkan regresor untuk mencegah peningkatan varians prediksi yˆ. Keseimbangan antara tujuan ini disebut memilih persamaan regresi "terbaik". Sayangnya, tidak ada definisi "terbaik" yang unik, dan algoritma yang berbeda dapat mengusulkan subset regresor kandidat yang berbeda sebagai yang terbaik.

Masalah pemilihan variabel sering dibahas dalam pengaturan yang diidealkan, mengasumsikan pengetahuan spesifikasi fungsional yang benar dari regresor dan ketiadaan outlier atau pengamatan berpengaruh. Dalam praktiknya, asumsi-asumsi ini jarang

terpenuhi. Analisis residu berguna untuk mengungkap bentuk fungsional, mengidentifikasi regresor kandidat baru, dan mendeteksi cacat dalam data, seperti outlier. Pendekatan iteratif sering digunakan, di mana strategi pemilihan variabel digunakan, dan model subset yang dihasilkan diperiksa untuk spesifikasi yang benar, outlier, dan pengamatan berpengaruh. Proses iteratif ini mungkin memerlukan pengulangan untuk menghasilkan model yang memadai.

Tidak satu pun dari prosedur pemilihan variabel menjamin persamaan regresi terbaik untuk satu set data tertentu. Biasanya, ada beberapa persamaan yang sama baiknya. Karena algoritma pemilihan variabel sangat bergantung pada komputer, analis sebaiknya menghindari terlalu bergantung pada hasil dari suatu prosedur tertentu.

10.1.2 *CONSEQUENCES OF MODEL MISSPECIFICATION* / DAMPAK DARI KESALAHAN SPESIFIKASI MODEL

Untuk memberikan motivasi untuk pemilihan variabel secara singkat meninjau konsekuensi dari spesifikasi model yang tidak benar. Dalam konteks ini, asumsuikan bahwa ada K regresor kandidat yaitu $x_1, x_2, \dots, x_K$ dengan $n \geq K + 1$ observasi pada regresor-regresor dan respons y. Model lengkap yang mencakup semua K regresor, yaitu

$$y_i = \beta_0 + \sum_{j=1}^{K} \beta_j x_{ij} + \varepsilon_i, \qquad i = 1,2, \dots, n$$

Atau

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

Diasumsikan bahwa daftar regresor kandidat mencakup semua variabel penting. Lalu, diasumsikan bahwa semua persamaan mencakup istilah intercept. Biarkan r menjadi jumlah regresor yang dihapus dari persamaan (10.1). Maka jumlah variabel yang tetap adalah $p = K + 1 - r$. Karena intercept disertakan, model subset berisi $p - 1 = K - r$. Model (10.1) dapat dituliskan sebagai

$$\boldsymbol{y} = \boldsymbol{X_p\beta_p} + \boldsymbol{X_r\beta_r} + \boldsymbol{\varepsilon}$$

Matriks X telah dipartisi menjadi Xp, matriks $(n \times p)$yang mencakup intercept dan $p - 1$ regresor untuk model subset, serta Xr, matriks $(n \times r)$ yang berisi regresor dihapus dari model. β dipartisi menjadi β_p dan β_r. Estimasi kuadrat terkecil untuk model lengkap adalah.

$$\widehat{\boldsymbol{\beta}}^* = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$$

Dan perkiraan residual varians σ^2 adalah

$$\hat{\sigma}^2 = \frac{y'y - \hat{\beta}^{*\prime}X'y}{n - K - 1}$$

Dengan nilai y, yaitu

$$y = X_p\beta_p + \varepsilon$$

Estimasi kuadrat terkecil untuk β_p adalah

$$\hat{\beta}_p = \left(X_p'X_p\right)^{-1}X_p'y$$

Estimasi varians residual adalah

$$\hat{\sigma}^2 = \frac{y'\left[I - X_p\left(X_p'X_p\right)^{-1}X_p'\right]y}{n - p}$$

Hasilya dapat diringkas sebagai berikut:

1. Ekspektasi dari estimasi kuadrat terkecil untuk $\hat{\beta}_p$ adalah

   $$E\left(\hat{\beta}_p\right) = \beta_p + \left(X_p'X_p\right)^{-1}X_p'X_r\beta_r = \beta_p + A\beta_r$$

   melibatkan unsur bias yang tergantung pada koefisien regresi dihapus (β_r) dan matriks alias A. Oleh karena itu, $\hat{\beta}_p$ adalah estimasi yang bias dari β_p kecuali jika koefisien regresi yang sesuai dengan variabel yang dihapus adalah nol atau variabel yang dipertahankan orthogonal terhadap variabel yang dihapus ($X_p'X_r = 0$).

2. Varians dari $\widehat{\boldsymbol{\beta}}_{\boldsymbol{p}}$ dan $\widehat{\boldsymbol{\beta}}^*$ adalah $\mathbf{Var}\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{p}}\right) = \widehat{\boldsymbol{\sigma}}^{\mathbf{2}}\left(\boldsymbol{X}_{\boldsymbol{p}}'\boldsymbol{X}_{\boldsymbol{p}}\right)^{-\mathbf{1}}$dan $\mathbf{Var}\ \left(\widehat{\boldsymbol{\beta}}^*\right) = \widehat{\boldsymbol{\sigma}}^{\mathbf{2}}(\boldsymbol{X}'\boldsymbol{X})^{-\mathbf{1}}$ dihitung dengan matriks invers produk dalam persamaan, dan matriks $\boldsymbol{Var}\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{p}}^*\right) - (\ \widehat{\boldsymbol{\beta}}^*)$ adalah positif semidefini, menunjukkan bahwa menghapus variabel tidak meningkatkan varian perkiraan parameter yang tersisa.

3. Karena $\hat{\beta}_p$ adalah estimasi yang bias dari β_p dan $\hat{\beta}_p^*$ tidak, perbandingan presisi estimasi parameter dari model penuh dan model subset diukur dalam kesalahan kuadrat rata-rata (MSE). MSE dari $\hat{\beta}_p$ adalah $\boldsymbol{\sigma}^{\mathbf{2}}\left(\boldsymbol{X}_{\boldsymbol{p}}'\boldsymbol{X}_{\boldsymbol{p}}\right)^{-\mathbf{1}} + \boldsymbol{A}\boldsymbol{\beta}_{\boldsymbol{r}}\boldsymbol{\beta}_{\boldsymbol{r}}'\boldsymbol{A}'$. Jika matriks $Var\left(\hat{\beta}_p^*\right) - MSE(\ \hat{\beta}_p)$ positif semidefini, ini menunjukkan bahwa estimasi kuadrat terkecil dari parameter dalam model subset memiliki kesalahan kuadrat rata-rata yang lebih kecil.

4. Estimasi parameter $\hat{\sigma}_*^2$ dari model adalah estimasi yang tidak bias untuk σ^2. Model subset

$$\boldsymbol{E(\hat{\sigma}^2) = \sigma^2 + \frac{B_r' C_r' \left[I - X_p (X_p' X_p)^{-1} X_p' \right] X_r \beta_r}{n - p}}$$

5. Misal, kita ingin memprediksi respon pada titik $x' = [x_p', x_r']$. Maka, untuk menghitung varians prediksi adalah

$$\boldsymbol{Var(\hat{y}^*) = \sigma^2 [1 + x'(X'X)^{-1} x]}$$

Dan prediksi Mean Square error

$$MSE(\hat{y}) = \sigma^2 \left[1 + x_p' (X_p' X_p)^{-1} x_p \right] + \left(x_p' A \beta_r - x_r' \beta_r \right)^2$$

Dalam mean square error, kita dapat menunjukkan bahwa

$$Var(\hat{y}^*) \geq MSE(\hat{y})$$

Asalkan matriks $Var(\hat{\beta}_p^*) - \beta_r \beta_r'$ adalah positif semidefinite.

Dengan menghapus variabel dari model, kita dapat meningkatkan presisi perkiraan parameter variabel yang tetap, meskipun beberapa variabel yang dihapus tidak signifikan. Ini juga berlaku untuk varians respons yang diprediksi. Meskipun penghapusan variabel berpotensi memasukkan bias ke dalam perkiraan koefisien variabel yang tetap dan respons, jika variabel yang dihapus memiliki efek kecil, MSE dari perkiraan yang bias akan lebih kecil dari varians dari perkiraan yang tidak bias. Artinya, manfaat mengurangi varians melebihi dampak bias yang mungkin terjadi. Namun, ada risiko dalam mempertahankan variabel yang diabaikan, terutama jika variabel tersebut memiliki koefisien nol atau koefisien yang jauh lebih kecil dari kesalahan standar yang sesuai dalam model lengkap. Risiko ini dapat meningkatkan varians perkiraan parameter dan respons yang diprediksi. Terakhir, dalam konteks penggunaan data retrospektif, seringkali terdapat cacat data seperti nilai terjauh, titik "ganjil", dan inkonsistensi akibat perubahan dalam sistem pengumpulan data dan pengolahan informasi organisasi. Cacat data ini dapat mempengaruhi pemilihan variabel dan menyebabkan misspecification model, yang mungkin hanya dapat diatasi oleh pengetahuan nonstatistik pembangun model tentang lingkungan masalah. Jika variabel yang dianggap penting memiliki rentang yang sangat terbatas karena pengendalian ketat, mungkin diperlukan pengumpulan data baru untuk upaya pembangunan model, dan desain eksperimen yang baik dapat membantu dalam hal ini.

10.1.3 *CRITERIA FOR EVALUATING SUBSET REGRESSION MODELS* / KRITERIA UNTUK MENILAI MODEL REGRESI SUBSET

Koefisien Deteriminasi Berganda Misalkan R^2 menunjukkan koefisien determinasi berganda untuk model regresi subset degan p suku, yaitu $p-1$ regressor dan suku intersep β_0.

$$\boldsymbol{R_p^2 = \frac{SS_R(p)}{SS_T} = 1 - \frac{SS_{Res}(p)}{SS_T}} \quad (10.8)$$

Di mana $SS_R(p)$ dan $SS_{Res}(p)$ menunjukkan jumlah kuadrat regresi dan jumlah kuadrat residu untuk model subset dengan p variabel. Perhatikan bahwa terdapat $\binom{K}{p-1}$ nilai R_p^2 untuk setiap nilai p, satu untuk setiap model subset yang mungkin dengan ukuran p. R_p^2 meningkat seiring dengan peningkatan p dan mencapai maksimum ketika $p = K+1$. Pendekatan umum ini diilustrasikan dalam Gambar 10.1, yang menunjukkan plot hipotetis nilai maksimum R_p^2untuk setiap subset berukuran p terhadap p. Biasanya, seorang analis memeriksa tampilan seperti ini dan kemudian menentukan jumlah regresor untuk model akhir pada titik di mana "lutut" dalam kurva menjadi jelas.
Karena kita tidak dapat menemukan nilai "optimal" R^2 untuk model regresi subset, kita harus mencari nilai "memuaskan". Aitkin [1974] mengusulkan satu solusi untuk masalah ini dengan menyediakan uji di mana semua model regresi subset yang memiliki R^2 tidak signifikan berbeda dari R^2untuk model penuh dapat diidentifikasi. Misalkan

$$R_0^2 = 1 - (1 - R_{K+1}^2)(1 + d_{a,n,K}) \quad (10.9)$$

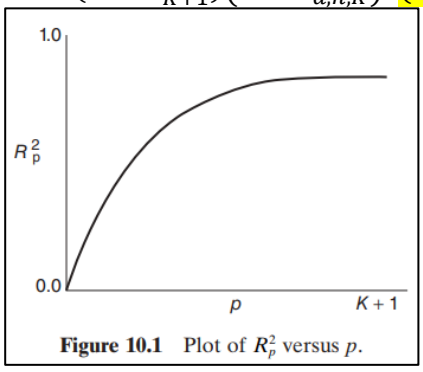


Figure 10.1 Plot of R_p^2 versus p.

Nilai R_{K+1}^2 menandakan model penuh. Subset R^2 adekuat α didefinisikan oleh Aitkin sebagai subset variabel regresor yang memiliki R^2 lebih besar dari R_0^2. Meskipun R^2 tidak langsung digunakan sebagai kriteria untuk jumlah regresor, R_p^2 digunakan untuk membandingkan $\binom{K}{p-1}$ model subset.

Adjusted R^2 Untuk menghindari kesulitan dalam menginterpretasi R^2. Statistik adjusted $R_{Adj,p}^2$memberikan alternatif yang lebih baik untuk R^2. Untuk jumlah variabel tetap p,

$$R_{adj,p}^2 = 1 - \frac{1}{n-p-1}\left(\frac{SSR(p)}{1-R_p^2}\right) \quad (10.10)$$

Jika, s regressor ditambahkan ke dalam model, $R^2_{Adj,p+8}$ akan melebihi $R^2_{Adj,p}$ jika dan hanya jika statistik F parsial untuk menguji signifikasin dari s regressor tambahan melebihi 1. Model dengan $R^2_{adj,p}$ maksimum dianggap sebagai pilihan optimum.

Residual Mean Square Mean square residual untuk model regresi subset, misalnya

$$MS_{Res}(p) = \frac{SS_{Res}(p)}{n-p} \quad (10.11)$$

dapat digunakan sebagai kriteria evaluasi model. Penurunan SS_{Res} seiring penambahan regresor menunjukkan peningkatan presisi model subset.

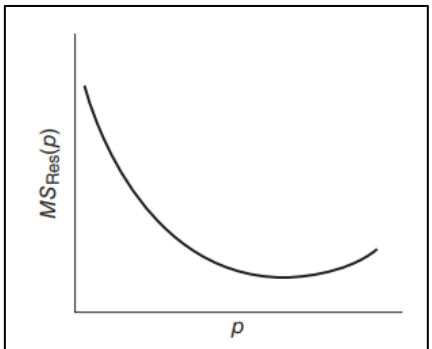


Figure 10.2 Plot of $MS_{Res}(p)$ versus p.

Ketika p meningkat, $MS_{Res}(p)$ pada awalnya menurun, kemudian stabil, dan akhirnya mungkin meningkat. Peningkatan akhir dalam $MS_{Res}(p)$ terjadi ketika pengurangan $SS_{Res}(p)$ dari penambahan regresor ke model tidak mencukupi untuk mengimbangi kehilangan satu derajat kebebasan dalam penyebut persamaan (10.11). Para pendukung kriteria MSRes(p) akan memplot MSRes(p) versus p dan memilih p berdasarkan:

1. Minimum $MS_{Res}(p)$
2. Nilai p sehingga $MS_{Res}(p)$ hampir sama dengan MS_{Res} untuk model penuh
3. Nilai p dekat dengan titik di mana $MS_{Res}(p)$ terkecil berubah ke atas

Model regresi subset yang meminimalkan $MS_{Res}(p)$ juga akan memaksimalkan $R^2_{adj,p}$. Perhatikan bahwa kedua kriteria minimum MSRes(p) dan maksimum adjusted \(R^2\) bersifat ekuivalen.

$$R^2_{adj.p} = 1 - \frac{MS_{Res}(p)}{SS_T/(n-1)}$$

Mallows's Cp Statistic Mallows mengusulkan statistik yang terkait dengan mean square error dari nilai yang diprediksi, yaitu

$$E[\hat{y}_i - E(y_i)]^2 = [E(y_i) - E(\hat{y}_i)]^2 + Var(\hat{y}_i) \quad (10.12)$$

Catatan: bahwa $E(y_i)$ adalah respons yang diharapkan dari persamaan regresi sebenarnya dan $E(\hat{y}_i)$ adalah respons yang diharapkan dari model subset dengan p variabel. Oleh karena itu, $E(y_i) - E(\hat{y}_i)$ adalah bias pada titik data ke-i. Sebagai konsekuensinya, dua istilah di sebelah kanan Eq. (10.12) adalah komponen squared bias dan variance dari mean square error. Definisikan total squared bias untuk persamaan dengan p variabel sebagai

$$SS_B(p) = \sum_{i=1}^{n} [E(y_i) - E(\hat{y}_i)]^2$$

Mendefinisikan kesalahan kuadrat rata-rata total terstandarisasi sebagai

$$r = \frac{SS_B(p)}{\sigma^2} + \frac{1}{\sigma^2} \sum_{i=1}^{n} var(\hat{y}_i) \quad (10.13)$$

Dapat ditunjukkan bahwa

$$\sum_{i=1}^{n} Var(\hat{y}_i) = p\sigma^2$$

Mensubstitusi $\sum_{i=1}^{n} Var(y_i)$ dan $SS_B(p)$ pada persamaan (10.13).

$$\Gamma_p = \frac{E[SS_{Res}(p)]}{\sigma^2} - n + 2p \quad (10.14)$$

Misalkan σ^2 adalah taksiran untuk σ^2. Kemudian mengganti $E[SS_{Res}(p)]$ dengan nilai observasi $SS_{Res}(p)$ menghasilkan taksiran Γ_p, maka

$$C_p = \frac{SS_{Res}(p)}{\hat{\sigma}^2} - n + 2p \quad (10.15)$$

Jika model p-term memiliki bias yang dapat diabaikan, maka $SS_B(p) = 0$. Akibatnya, $E[SS_{Res}(p)] = (n - p)\sigma^2$.

Dalam menggunakan kriteria C_p, visualisasi plot C_p, terhadap p sangat membantu sebagai fungsi dari p. Persamaan regresi dengan sedikit bias memiliki nilai C_p yang dekat dengan garis $C_p = p$, sementara persamaan dengan bias signifikan berada di atas garis tersebut. C_p kecil umumnya diinginkan. Perhitungan C_p membutuhkan perkiraan σ^2 yang tidak bias. Meskipun sering kali menggunakan $MS_{Res}(K + 1)$ sebagai alternatif, harus digunakan perkiraan yang baik dari σ^2. Visualisasi plot membantu menentukan model yang

paling sesuai, dengan mempertimbangkan trade-off antara bias dan kesalahan prediksi rata-rata.

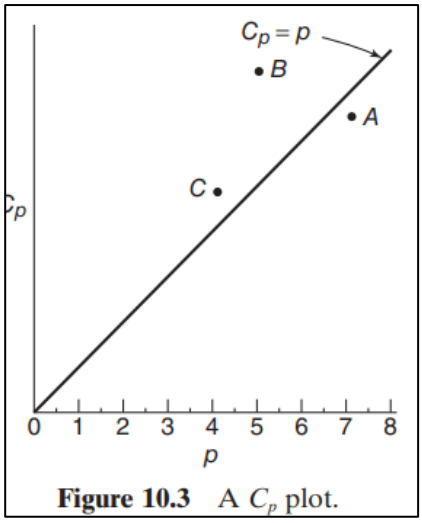


Figure 10.3 A C_p plot.

Kriteria Informasi Akaike dan Bayesian Analog (BIC) Akaike mengusulkan AIC sebagai kriteria informasi berdasarkan maksimalkan entropi yang diharapkan dari model. Pada dasarnya, AIC adalah ukuran log-likelihood yang dihukum. AIC adalah

$$AIC = -2\ln(L) + 2\,p,$$

Di mana p adalah jumlah parameter dalam model. Dalam kasus regresi least-square,

$$AIC = n\ln\left(\frac{SS_{Res}}{n}\right) + 2\,p.$$

AIC memiliki persamaan yang mirip dengan RAdj^2 dan Mallows Cp. Saat menambahkan regresor ke dalam model, SSRes tidak dapat meningkat. Ada beberapa ekstensi Bayesian dari AIC. Schwartz (1978) dan Sawa (1978) adalah dua yang lebih populer. Keduanya disebut BIC (Bayesian information criterion). Kriteria Schwartz (BIC Sch) adalah

$$BIC_{Sch} == 2\ln(L) + p\ln(n)$$

Kriteria ini memberikan hukuman yang lebih besar untuk menambahkan regresor saat ukuran sampel meningkat. Untuk regresi least-square, kriteria ini adalah

$$BIC_{Sch} = n\ln\left(\frac{SS_{Res}}{n}\right) + p\ln(n)$$

R menggunakan kriteria ini sebagai BIC-nya. SAS menggunakan kriteria Sawa, yang melibatkan istilah hukuman yang lebih rumit, melibatkan σ^2 dan σ^4, yang diestimasi oleh MSRes dari model penuh.

Kriteria AIC dan BIC semakin populer dan sering digunakan dalam prosedur pemilihan model yang melibatkan situasi pemodelan yang lebih rumit daripada regresi kuadrat terkecil biasa, seperti situasi model campuran. Kriteria ini umumnya digunakan dalam model-model linear umum (Bab 13).

Penggunaan Regresi dan Kriteria Evaluasi Model Terdapat beberapa kriteria yang dapat digunakan untuk mengevaluasi model regresi subset. Pemilihan kriteria model seharusnya terkait dengan penggunaan yang dimaksudkan dari model tersebut. Beberapa penggunaan potensial dari regresi meliputi (1) deskripsi data, (2) prediksi dan estimasi, (3) estimasi parameter, dan (4) kontrol.

Jika tujuannya adalah untuk mendapatkan deskripsi yang baik dari suatu proses atau memodelkan sistem yang kompleks, pencarian persamaan regresi dengan jumlah kuadrat sisa kecil diperlukan. Kriteria umumnya adalah mengurangi variabel SS_{Res} jika hanya menghasilkan peningkatan kecil dalam jumlah kuadrat sisa. Secara umum, kita ingin mendeskripsikan sistem dengan sedikit regresor sebanyak mungkin sambil menjelaskan sebagian besar variasi dalam y.

Seringkali, persamaan regresi digunakan untuk prediksi observasi masa depan atau estimasi respons rata-rata. Umumnya, kita ingin memilih regresor sehingga kesalahan kuadrat rata-rata prediksi diminimalkan. Statistik PRESS, yang diperkenalkan dalam Bab 4, bisa digunakan untuk mengevaluasi persamaan kandidat yang dihasilkan oleh suatu prosedur pembangkitan subset. Model regresi dengan p-term

$$PRESS_p = \sum_{i=1}^{n} [y_i - \hat{y}_i]^2 = \sum_{i=1}^{n} \left(\frac{e_i}{i - h_{ii}} \right)^2$$

Jika tujuannya adalah estimasi parameter, maka perlu mempertimbangkan baik bias yang muncul dari penghapusan variabel maupun varians koefisien yang diestimasi. Ketika regresor sangat multikolinear, estimasi kuadrat terkecil dari koefisien regresi individual dapat sangat buruk.

Ketika model regresi digunakan untuk kontrol, estimasi parameter yang paling penting. Ini mengimplikasikan bahwa standar kesalahan koefisien regresi seharusnya kecil. Selain itu, karena penyesuaian yang dilakukan pada x untuk mengontrol y akan sebanding dengan $\hat{\beta}'s$, koefisien regresi seharusnya secara akurat merepresentasikan efek regresor.

Jika regresor sangat multikolinear, $\hat{\beta}'s$ mungkin merupakan estimasi yang sangat buruk dari efek regresor individual.

10.2 *COMPUTATIONAL TECHNIQUES FOR VARIABLE SELECTION* / TEKNIK KOMPUTASI UNTUK PEMILIHAN VARIABEL

Setelah diketahui bahwa mempertimbangkan model regresi dengan menggunakan sebagian dari *regressor* kandidat merupakan salah satu hal yang sangat penting. Selanjutnya, pada bagian ini akan dilakukan pembahasan mengenai teknik komputasi untuk menghasilkan model regresi subset dan mengilustrasikan kriteris untuk evaluasi dalam masing-masing model.

10.2.1 *ALL POSSIBLE REGRESSIONS* / SEMUA KEMUNGKINAN REGRESI

Dalam menggunakan prosedur ini, diharuskan menggunakan analisis untuk mencocokkan semua persamaan regresi yang melibatkan satu kandidat *regressor,* dua kandidat *regressor,* dan seterusnya. Persamaan-persamaan ini dievaluasi berdasarkan beberapa kriteria yang sesuai dengan model regresi yang dipilih. Jika dilakukan asumsi jika istilah dari *intersep* β_0 terdapat dalam semua persamaan, maka jika terdapat κ yang merupakan kandidat *regressor*, maka terdapat 2^{κ}. Dimana 2^{κ} merupakan jumlah total persamaan yang harus diestimasi dan diperiksa. Dengan contoh, jika terdapat $\kappa = 4,\ maka\ 2^4 = 16$. Terdapat 16 persamaan yang regresi yang mungkin untuk di periksa. Dapat diketahui bahwa jumlah persamaan regresi yang mungkin untuk diperiksa meningkat seiring dengan bertambahnya jumlah kandidat *regressor.*

Untuk penerapan dari algoritma yang efisien untuk semua regresi yang mungkin pada bagian pembahasan ini dilakukan penggunaan Minitab dan SAS yang diterapkan pada data semen hald dengan variabel x1, x2, x3, x4.

EXAMPLE / CONTOH 10.1 DATA SEMEN HALD

Dengan menggunakan studi kasus, seperti berikut :

Example 10.1 The Hald Cement Data

Hald [1952]† presents data concerning the heat evolved in calories per gram of cement (y) as a function of the amount of each of four ingredients in the mix: tricalcium aluminate (x_1), tricalcium silicate (x_2), tetracalcium alumino ferrite (x_3), and dicalcium silicate (x_4). The data are shown in Appendix Table B.21. These reflect quite serious problems with multicollinearity. The VIFs are:

x1: 38.496
x2: 254.423
x3: 46.868
x4: 282.513

We will use these data to illustrate the all-possible-regressions approach to variable selection.

Pada studi kasus diatas, HALD [1952] menyajikan data mengenai panas yang di hasilkan dalam kalori per-gram semen (y) sebagai fungsi dari jumlah masing-masing empat bahan dalam vampuran : tri-kalsium aluminat (x_1), trikalsium silikat (x_2), tetrakalsium alumino ferit (x_3), dan dikalsinasi silikat (x_4). Dengan menggunakan nilai pada tabel 2.1, seperti di bawah ini.

TABLE B.21 Hald Cement Data

Observation					
i	y_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	109.4	10	68	8	12

Dapat diperlihatkan bahwa terdapat masalah yang cukup serius dengan multikolinearitas. Dengan nilai VIF *(Variance Inflation Factor)* yang merupakan ukutan yang digunakan dalam analisis regresi untuk mengevaluasi sejauh mana multikolinearitas terjadi antara variabel-variabel independent dalam suatu model regresi.

$$x_1 = 38{,}496$$
$$x_2 = 254{,}423$$
$$x_3 = 46{,}868$$
$$x_4 = 282{,}513$$

Lalu dengan menggunakan data-data yang terdapat diatas, dilakukan ilustrasi dengan pendekatan *all-potential-regressions* / semua kemungkinan regresi untuk selesai yang dapat bervariasi. Dan diperoleh tabel *summary of all possible refressions for the Hald Cemet Data* / tabel ringkasan semua kemungkinan regresi untuk semua Data Semen Hald, seperti dibawah ini.

TABEL 10.1 *SUMMARY OF ALL POSSIBLE REGRESSIONS FOR THE HALD CEMENT DATA* / RINGKASAN SEMUA KEMUNGKINAN REGRESI UNTUK DATA SEMEN HALD

TABLE 10.1 Summary of All Possible Regressions for the Hald Cement Data

Number of Regressors in Model	p	Regressors in Model	$SS_{Res}(p)$	R_p^2	$R_{Adj,p}^2$	$MS_{Res}(p)$	C_p
None	1	None	2715.7635	0	0	226.3136	442.92
1	2	x_1	1265.6867	0.53395	0.49158	115.0624	202.55
1	2	x_2	906.3363	0.66627	0.63593	82.3942	142.49
1	2	x_3	1939.4005	0.28587	0.22095	176.3092	315.16
1	2	x_4	883.8669	0.67459	0.64495	80.3515	138.73
2	3	x_1x_2	57.9045	0.97868	0.97441	5.7904	2.68
2	3	x_1x_3	1227.0721	0.54817	0.45780	122.7073	198.10
2	3	x_1x_4	74.7621	0.97247	0.96697	7.4762	5.50
2	3	x_2x_3	415.4427	0.84703	0.81644	41.5443	62.44
2	3	x_2x_4	868.8801	0.68006	0.61607	86.8880	138.23
2	3	x_3x_4	175.7380	0.93529	0.92235	17.5738	22.37
3	4	$x_1x_2x_3$	48.1106	0.98228	0.97638	5.3456	3.04
3	4	$x_1x_2x_4$	47.9727	0.98234	0.97645	5.3303	3.02
3	4	$x_1x_3x_4$	50.8361	0.98128	0.97504	5.6485	3.50
3	4	$x_2x_3x_4$	73.8145	0.97282	0.96376	8.2017	7.34
4	5	$x_1x_2x_3x_4$	47.8636	0.98238	0.97356	5.9829	5.00

Karena terdapat nilai $\kappa = 4\ kandidat\ regressor,\ maka\ 2^4 = 16$ persamaan yang selalu menyertakan nilai *intersep* β_0. Dimana hasil fitting dari 16 persamaan ini telah di tampilkan pada tabel diatas, yang berisi nilai statistic $R^2{}_p, R^2{}_{adj,p}, MS_{Res}(p), dan\ C_p$.

Setelah mendapatkan nilai-nilai yang terdapat pada tabel tersebut selanjutnya, bisa diperlihatkan nilai *least-squares estimates for all possible regressions (Hald Cemet Data)* / estimasi kuadrat terkecil untuk semua kemungkinan regresi (Data Semen Hald), seperti dibawah ini.

TABEL 10.2 *LEAST-SQUARES ESTIMATES FOR ALL POSSIBLE REGRESSIONS (HALD CEMET DATA)* / ESTIMASI KUADRAT TERKECIL UNTUK SEMUA KEMUNGKINAN REGRESI (DATA SEMEN HALD)

TABLE 10.2 Least-Squares Estimates for All Possible Regressions (Hald Cement Data)

Variables in Model	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
x_1	81.479	1.869			
x_2	57.424		0.789		
x_3	110.203			−1.256	
x_4	117.568				−0.738
x_1x_2	52.577	1.468	0.662		
x_1x_3	72.349	2.312		0.494	
x_1x_4	103.097	1.440			−0.614
x_2x_3	72.075		0.731	−1.008	
x_2x_4	94.160		0.311		−0.457
x_3x_4	131.282			−1.200	−0.724
$x_1x_2x_3$	48.194	1.696	0.657	0.250	
$x_1x_2x_4$	71.648	1.452	0.416		−0.237
$x_2x_3x_4$	203.642		−0.923	−1.448	−1.557
$x_1x_3x_4$	111.684	1.052		−0.410	−0.643
$x_1x_2x_3x_4$	62.405	1.551	0.510	0.102	−0.144

Setelah dapat menampilkan nilai estimasi kuadrat terkecil dari koefisien regresi, untuk sifat parsial dari koefisien regresi dapat di perlihatkan dengan gambar plot pemeriksaan, seperti dibawah ini.

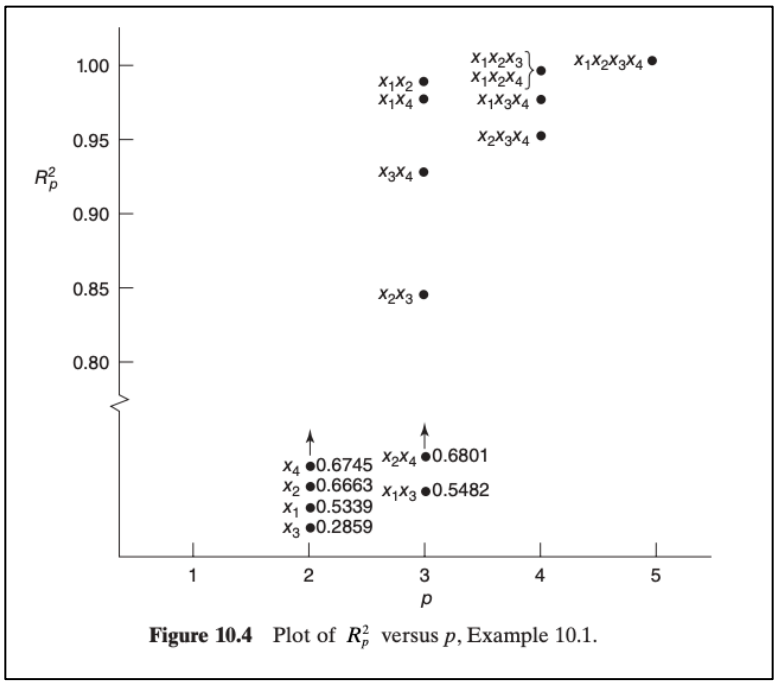


Figure 10.4 Plot of R_p^2 versus p, Example 10.1.

GAMBAR 10.4

Dengan contoh, Ketika mempertimbangkan x_2, maka terdapat nilai estimasi kuadrat terkecil dari efek x_2, adalah 0,789. Dan jika x_4 ditambahkan ke dalam model, maka efek $x_2$mengalami perubahan berkurang lebih dari 50% menjadi 0,311. Penambahan lebih lanjut dari x_3 dapat mengubah efek x_2 menjadi -0,923. Dimana dari nilai-nilai ini dapat diperlihatkan bahwa estimasi kuadrat terkecil dari koefisien regresi individu sangat bergantung pada *regressor* lain dalam model. Pada data semen hal konsisten terdapat masalah serius dengan multikolinieritas,

Dengan melakukan pertimbangan evaluasi model subset kriteria ${R^2}_p$ yang digambarkan pada plot gambar diatas, dapat disimpulkan bahwa setelah terdapat dua *regressor* dimasukkan kedalam model, peningkatan R^2 tidak signifikan dengan memasukkan variabel tambahan. Kedua model regresi (x_1, x_2) dengan (x_1, x_4) pada dasarnya memiliki nilai R^2 yang sama. Dari segi kriteria ini, tidak ada perbedaan anatara kedua model dalam pemilihan persamaan regresi akhir. Namun, lebih disarankan untuk menggunakan model (x_1, x_4), karena x_4 memberikan model satu regresi yang terbaik. Dengan menggunakan persamaan 10.9 dengan menggunakan $\alpha = 0{,}05$. Diperoleh perhitungan nilai R^2, seperti dibawah ini.

$$ {R^2}_0 = 1 - (1 - {R^2}_5) \left(1 + \frac{4F_{0.05,4,8}}{8}\right) $$

$$= 1 - 0.01762\left[1 + \frac{4(3.84)}{8}\right] = 0.94855.$$

Dari perhitungan tersebut, setiap model regresi subset dengan nilai $R^2 > R^2 = 0{,}94855$ dianggap memadai pada tingkat signifikasi 0,05. Hal ini mengindikasi bahwa nilai R^2 pada model tersebut tidak berbeda secara signifikan dengan ${}_{k+1}^{0}R^2$.

TABEL 10.3 ***MATRIX OF SIMPLE CORRELATIONS FOR HALD'S DATA IN EXAMPLE*** **10.1 / MATRIKS KORELASI SEDERHANA UNTUK DATA HALD DALAM CONTOH 10.1**

TABLE 10.3 Matrix of Simple Correlations for Hald's Data in Example 10.1

	x_1	x_2	x_3	x_4	y
x_1	1.0				
x_2	0.229	1.0			
x_3	–0.824	–0.139	1.0		
x_4	–0.245	–0.973	0.030	1.0	
y	0.731	0.816	–0.535	–0.821	1.0

Pada tabel diatas, dilakukan proses pemeriksaan korelasi berpasangan sederhana antara x_i dan x_j dan antara x_i dan y. Dapat diperlihatkan bahwa pasangan *regressor* antara (x_1, x_3) dan (x_2, x_4) sangat berkorelasi. Hal ini dikarenakan.

$$r_{13} = -0.824 \; dan \; r_{24} = -0.973.$$

Dalam analisis regresi, penambahan variabel tambahan dalam suatu model tidak selalu memberikan manfaat yang signifikan jika informasi yang dimilikinya sudah tercakup dalam variabel yang sudah ada dalam model. Pada gambar plot $MS_{Res}\ (p)\ versus\ (p)$ yang akan diperlihatkan pada gambar 10.5, seperti dibawah.

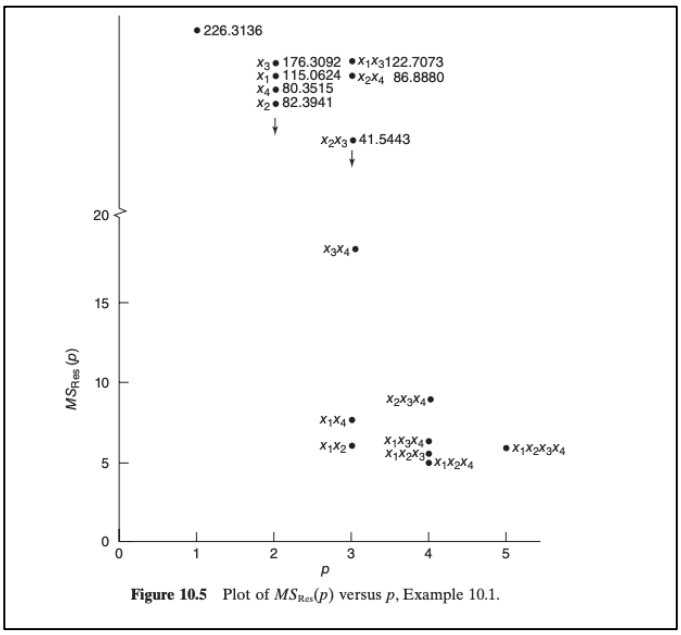
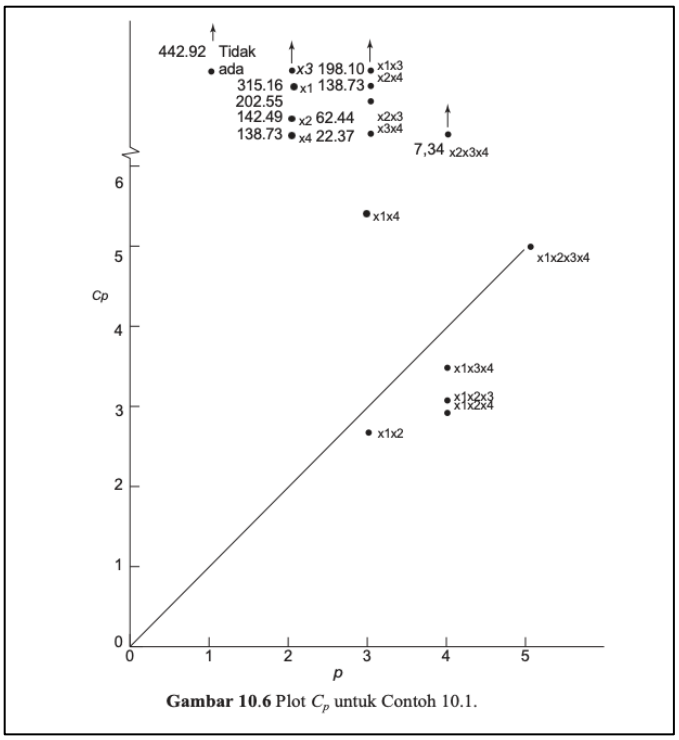


Figure 10.5 Plot of $MS_{Res}(p)$ versus p, Example 10.1.

GAMBAR 10.5

Dengan model kuadrat tengah residual minimum adalah (x_1, x_2, x_4), dengan nilai $MS_{Res}\ (4) = 5{,}3303$. Dapat diperhatikan bahwa model ini meminimumkan nilai $MS_{Res}\ (p)$ dan memaksimumkan nilai R^2yang telah disesuaikan. Akan tetapi, dua dari model regresi lainnya, yakni $[(x_1, x_2, x_3)$ dan $(x_1, x_3, x_4)]$ dan model dua regresi $[(x_1, x_2)$ dan $(x_1, x_4)]$ dapat memiliki nilai residual *mean square* yang sebanding, jika salah satu dari (x_1, x_2) atau (x_1, x_4) terdapat dalam model. Dengan nilai model subset (x_1, x_2) mungkin lebih tepat daripada (x_1, x_4). Hal ini dikarenakan memiliki nilai yang lebih kecil dari rata-rata kuadrat residual.

Untuk gambar plot C_p yang akan diperlihatkan pada gambar 10.6, seperti dibawah.



Gambar 10.6 Plot C_p untuk Contoh 10.1.

GAMBAR 10.6

Untuk bisa melakukan ilustrasi perhitungan, dapat dipermisalkan melakukan pengambilan nilai $\hat{\sigma} = 5{,}9829$ $(MS_{Res}\ dari\ model\ lengkap)$ dan menghitung C_3 untuk model (x_1, x_4). Dengan menggunakan persamaan 10.15. Diperoleh perhitungan C_3, seperti dibawah ini.

$$C_3 = \frac{SSRes\ (3)}{\sigma^2} - n + 2p = \frac{74{,}7621}{5{,}9829} - 13 + 2(3) = 5{,}50.$$

Dari pemeriksaan terhadap plot tersebut, dapat ditemukan terdapat empat model yang dapat diterima, diantaranya yakni (x_1, x_2), (x_1, x_2, x_3), dan (x_1, x_2, x_4), dan (x_1, x_3, x_4). Dengan mempertimbangkan nilai C yang lebih kecil, model yang paling sederhana (x_1, x_2) mungkin menjadi pilihan yang tepat. Namun, penting untuk diingat bahwa tidak terdapat pilihan yang jelas mengenai persamaan regresi terbaik. Dapat diambil contoh, bahwa persamaan C_p minimum adalah (x_1, x_2) dan persamaan MS_{Res} dari persamaan tersebut adalah (x_1, x_2, x_4). Semua model kandidat "final" harus dilakukan pengujian secukupnya, termasuk penyelidikan titik pengungktit, pengaruh, dan multikolinieritas. Hal ini dapat diilustrasikan pada tabel 10.4 yang menguji dua model, yakni (x_1, x_2) dan (x_2, x_4) sehubungan dengan PRESS dan faktor inflasi varians (VIF), dimana kedua model tersebut memiliki nilai PRESS yang mirip (yakni dua kali jumlah kuadrat residual untuk persamaan MS_{Res} minimum) dan nilai R^2 untuk prediksi yang dihitung dari PRESS sangat serupa untuk kedua model.

Namun, untuk $x_2\ dan\ x_4$ sangat multikoliniear, hal ini dibuktikan oleh faktor inflasi varians yang lebih besar dari (x_1, x_2, x_4).Dikarenakan model (x_1, x_2)

dan (x_2, x_4) memilik nilai PRESS yang sama, maka lebih direkomendasikan menggunakan model (x_1, x_2) berdasarkan minimnya multikolinieritas pada model.

Pembangkitan Efisien Dari Semua Kemungkinan Regresi

Terdapat beberapa algoritma yang dipergunakan untuk semua kemungkinan regresi diantaranya, yakni Furnival [1971], Furnival dan Wilson [1974], Gartside [1965, 1971], Morgan dan Tatar [1972], dan Schatzoff, Tzao, dan Fienberg [1968] semua algoritma ini menggunakan dasar untuk melakukan perhitungan 2^k model subset sehingga model subset yang berurutan hanya berbeda satu variabel dengan perhitungan menggunakan metode numerik yang didasarkan reduksi Gauss-Jordan. Dalam pembahasa ini dipergunakan program komputer Minitab dan SAS yang di terapkan pada data semen hald dan outputnya dapat diperlihatkan pada gambar 10.7. Program ini memungkinkan didapatkan model regresi subset terbaik dari setiap ukuran untuk $1 \leq p \leq K+1$ dan dapat menampilkan $C_p, R^2, dan\ MS_{Res}\ (p)$. Dapat juga ditampilkan nilai statistik $C_p, {R^2}_P, {R^2}_{Adj.p}, dan\ S = \sqrt{MS_{Res}}\ (p)$ dalam beberapa model untuk setiap nilai *p*. Program ini memiliki kemampuan untuk identifikasi model regresi subset terbaik dengan $m \leq 5$. Prosedur regresi yang ada dapat sangat efisien dan dapat memproses sekitar 30 kandidat regressor dengan waktu komputasi yang sebanding dengan algoritma regresi tipe *stepwise* yang akan dibahas pada 10.2.2. Terdapat tabel perbandingan untuk memandingkan model antara (x_1, x_2) dan (x_1, x_2, x_4), seperti dibawah ini.

TABEL 10.4 ***COMPRASIONS OF TWO MODELS FOR HALD'S CEMENT DATA*** **/ PERBANDINGAN DUA MODEL UNTUK DATA SEMEN HALD**

TABLE 10.4 Comparisons of Two Models for Hald's Cement Data

Observation i	$\hat{y} = 52.58 + 1.468x_1 + 0.662x_2$[a]			$\hat{y} = 71.65 + 1.452x_1 + 0.416x_2 - 0.237x_4$[b]		
	e_i	h_{ii}	$[e_i/(1-h_{ii})]^2$	e_i	h_{ii}	$[e_i/(1-h_{ii})]^2$
1	−1.5740	0.25119	4.4184	0.0617	0.52058	0.0166
2	−1.0491	0.26189	2.0202	1.4327	0.27670	3.9235
3	−1.5147	0.11890	2.9553	−1.8910	0.13315	4.7588
4	−1.6585	0.24225	4.7905	−1.8016	0.24431	5.6837
5	−1.3925	0.08362	2.3091	0.2562	0.35733	0.1589
6	4.0475	0.11512	20.9221	3.8982	0.11737	19.5061
7	−1.3031	0.36180	4.1627	−1.4287	0.36341	5.0369
8	−2.0754	0.24119	7.4806	−3.0919	0.34522	22.2977
9	1.8245	0.17195	4.9404	1.2818	0.20881	2.6247
10	1.3625	0.55002	9.1683	0.3539	0.65244	1.0368
11	3.2643	0.18402	16.0037	2.0977	0.32105	9.5458
12	0.8628	0.19666	1.1535	1.0556	0.20040	1.7428
13	−2.8934	0.21420	13.5579	−2.2247	0.25923	9.0194
		PRESS x_1, x_2 = 93.8827			PRESS x_1, x_2, x_4 = 85.3516	

[a] $R^2_{\text{Prediction}} = 0.9654$, $\text{VIF}_1 = 1.05$, $\text{VIF}_2 = 1.06$.

[b] $R^2_{\text{Prediction}} = 0.9684$, $\text{VIF}_1 = 1.07$, $\text{VIF}_2 = 18.78$, $\text{VIF}_4 = 18.94$.

Dengan output perhitungan computer Minitab untuk algoritma regresi semua kemungkinan dari teori Furnival dan Wilson, seperti dibawah ini.

Best Subsets Regression: y versus x1, x2, x3, x4

Response is y

Vars	R-Sq	R-Sq(adj)	C-p	S	x 1	x 2	x 3	x 4
1	67.5	64.5	138.7	8.9639				X
1	66.6	63.6	142.5	9.0771		X		
1	53.4	49.2	202.5	10.727	X			
1	28.6	22.1	315.2	13.278			X	
2	97.9	97.4	2.7	2.4063	X	X		
2	97.2	96.7	5.5	2.7343	X			X
2	93.5	92.2	22.4	4.1921			X	X
2	84.7	81.6	62.4	6.4455		X	X	
2	68.0	61.6	138.2	9.3214		X		X
3	98.2	97.6	3.0	2.3087	X	X		X
3	98.2	97.6	3.0	2.3121	X	X	X	
3	98.1	97.5	3.5	2.3766	X		X	X
3	97.3	96.4	7.3	2.8638		X	X	X
4	98.2	97.4	5.0	2.4460	X	X	X	X

Figure 10.7 Computer output (Minitab) for Furnival and Wilson all-possible-regression algorithm.

GAMBAR 10.7

10.2.2 *STEPWISE REGRESSION METHODS /* METODE REGRESI BERTAHAP

Ketika melakkan evaluasi dari semua regresi mungkin dapat membebani secara komputasi, untuk mengatasi hal ini bisa dilakukan penerapan berbagai metode yang telah dikembangan untuk melakukan evaluasi pada hanya sejumlah kecil model regresi subset dengan menambahkan atau menghapus *regressor* satu persatu. Metode-metode ini umumnya disebut sebagai *stepwise-type procedures* atau prosedur tipe bertahap, yang dikelompokkan dalam tiga kategori besar, yakni *forward selection* / seleksi maju*, backward elimination* / eliminasi mundur*,* dan *stepwise regression* / regresi bertahap. Dimana pada bagian pembahasan ini akan dijelaskan dan melakukan ilustrasi dari prosedur-prosedur tersebut secara singkat.

Forward Selection **/ Seleksi Maju** dimana asumsi ini dimulai bahwa tidak adanya *regressor* dalam model selain *intersep*. Dilakukan proses penemuan subset yang optimal dengan memasukkan *regressor* kedalam model satu persatu. *Regressor* pertama yang dipilih untuk dimasukkan kedalam persamaan adalah *regressor* yang memiliki korelasi sederhana terbesar dengan variabel respon y. Dipermisalkan *regressor* tersebut adalah x_1 yang merupakan *regressor* dengan nilai statistic F terbesar untuk uji signifikasi regresi, *regressor* ini akan dimasukkan jika F statistik melebihi nilai F yang dipilih sebelumnya (FIN). *Regressor* kedua yang dipilih untuk dimasukkan adalah *regressor* yang memiliki korelasi terbesar dengan y setelah disesuaikan dengan pengaruh *regressor* pertama yang dimasukkan, yakni x_1 dan disebut sebagai korelasi parsial. Dengan korelasi-korelasi tersebut merupakan korelasi sederhana antara residual dari regresi $\hat{y} = \widehat{\beta_0} + \widehat{\beta_{1x1}}$ dengan residual dari regresi kandidat *regressor* lainnya pada x_1 dipermisalkan dengan persamaan $\widehat{x_j} = \widehat{\alpha_{0j}} + \hat{\alpha}_{1jx1}$ dengan j = 2,3,…,k.

Dipermisalkan pada langkah 2, *regressor* dengan korelasi parsial tertinggi dengan y adalah x_2 dan dapat diimplementasikan bahwa statistik F parsial terbesar, yakni.

$$F = \frac{SS_R\ (x_2\,|x_1)}{MS_{Res}\ (x_1, x_2)}$$

Dengan syarat jika nilai $F > FIN$, maka x_2 ditambahkan kedalam model. Secara umum dapat dijelaskan bahwa pada setiap langkah dengan *regressor* yang memiliki korelasi parsial tertinggi dengan y (ekuivalen denga statistic F parsial terbesar dengan *regressor* lain yang sudah ada dalam model) akan ditambahkan kedalam model jika F statistik parsialnya melebihi *FIN* dan akan berhenti jika F statistik parsial tidak lebih dari FIN atau jika *regressor* kandidat terakhir ditambahkan dalam model. Dengan penjelasan bahwa prosedur computer dapat memasukkan atau menghapus variabel t, dikarenakan $t^2_{\frac{\alpha}{2}}, v = F_{\alpha,1,v}$. Untuk ilustrasi pada pembahasan ini dilakukan menggunakan Minitab dan SAS, dengan contoh soal seperti dibawah ini.

***EXAMPLE* / CONTOH 10.2 *FORWARD SELECTION – HALD CEMENT DATA* / SELEKSI MAJU DATA SEMEN HALD**

Dengan melakukan penerapan prosedur seleksi maju pada data semen hald yang diberikan dicontoh 10.1. Digambar 10.8 yang akan ditampilkan dibawah merupakan hasil yang diperoleh ketika algoritma seleksi maju data semen hald menggunakan Minitab dengan melakukan penetapan nilai batas untuk memasukkan variabel dengan tingkat kesalahan tipe I_α dan menggunakan statistik t untuk pengambilan keputusan terkait pemilihan variabel dengan korelasi parsial tersbesar dengan y di tambahkan kemodel ketika $|t| > t_{/2}$. Pada perhitungan ini digunakan $\alpha = 0{,}25$. Dan diperoleh hasil, seperti dibawah ini.

Stepwise Regression: y versus x1, x2, x3, x4

Forward selection. Alpha-to-enter: 0.25

Response is y on 4 predictors, with N = 13

Step	1	2	3
Constant	117.57	103.10	71.65
x4	-0.738	-0.614	-0.237
T-Value	-4.77	-12.62	-1.37
P-Value	0.001	0.000	0.205
x1		1.44	1.45
T-value		10.40	12.41
P-Value		0.000	0.000
x2			0.42
T-Value			2.24
R-Value			0.052
S	8.96	2.73	2.31
R-Sq	67.45	97.25	98.23
R-Sq(adj)	64.50	96.70	97.64
Mallows C-p	138.7	5.5	3.0

Figure 10.8 Forward selection results from Minitab for the Hald cement data.

GAMBAR 10.8

Dari tabel 10.3 dapat diketahui bahwa *regressor* yang paling berkorelasi tinggi dengan y adalah x_4 dengan ($r_{4y} = -0{,}821$) dan karena t statistik untuk model dengan x_4

adalah 4,77 dimana nilai ini melebihi $t_{0,25/2,11} = 1,21$, maka x_4 ditambahkan dalam persamaan. Untuk langkah kedua, x_1 ditambahkan karena memiliki korelasi parsial tertinggi dengan y memiliki nilai F parsial t statistik = 1040 > $t_{0,25/2,10}$ = 1,22. Hal tersebut juga berlaku untuk x_2 dengan nilai t statistik 2,24 > $t_{0,25/2,9}$ = 1,23, dan berhenti pada x_3 dikarenakan t statistik yang dihasilkan tidak melebihi $t_{0,25/2,8} = 1,24$. Sehingga, diperoleh model akhir.

$$\hat{y} = 71,6483 + 1,4519x_1 + 0,4161x_2 - 0,2365x_4.$$

***Backward Elimination* / Seleksi Mundur** dimana asumsi ini dimulai dengan menemukan model yang baik dengan memulai model yang mencakup semua k kandidat *regressor*. Kemudian, jika F parsial (setara dengan t statistik) dihitung untuk setiap *regressor* seolah-olah itu adalah variabel terakhir yang masuk kedalam model. Nilai terkecik dari F parsial atau t statistik ini dibandingkan dengan nilai yang dipilih sebelumnya, yakni F_{OUT} (atau t_{OUT}). Dimisalkan terdapat F parsial terkecil (atau t) dengan nilai lebih kecil dari F_{OUT} (atau t_{OUT}), maka *regressor* tersebut dikeluarkan dari model. Lalu, diperoleh model regresi dengan K-1 yang telah sesuai dan dilakukan perhitungan statistik parsial F (atau t) untuk model baru ini dihitung dan prosedur diulangi. Dapat dikatakan bahwa pada algoritma ini berhenti ketika F (atau t) parsial terkecil tidak kurang dari batas yang telah ditentukan sebelumnya pada F_{OUT} (atau t_{OUT}).

EXAMPLE / CONTOH 10.3 *BACKWARD ELIMINATION – HALD CEMENT DATA* / SELEKSI MUNDUR DATA SEMEN HALD

Dengan melakukan penerapan prosedur seleksi mundur pada data semen hald yang diberikan dicontoh 10.1. Pada gambar 10.9 akan disajikan hasil dari pengunaan eliminasi mundur menggunakan Minitab dengan inisiasi nilai $\alpha = 0,10$. Dengan melakukan penerapan algoritma menggunakan t statistik untuk menghilangkan variabel dengan syarat *regressor* akan dihilangkan jika nilai absolute t statistik < $t_{OUT} = t_{\frac{0,1}{2},n-p.}$ Dengan penerapan, pada langkah 1 diperoleh hasil fitting model penuh dengan nilai t terkecil sebesar 0,14 dan berhubungan dengan x_3. Maka, karena t = 0,14 < $t_{OUT} = t_{\frac{0,10}{2},8} = 1,86$, maka x_3 dikeluarkan dari model. Dan diperoleh hasil fitting, seperti dibawah ini.

Stepwise Regression: y versus x1, x2, x3, x4

Backward elimination. Alpha-to-Remove: 0.1

Response is y on 4 predictors, with N = 13

Step	1	2	3
Constant	62.41	71.65	52.58
x1	1.55	1.45	1.47
T-Value	2.08	12.41	12.10
P-Value	0.071	0.000	0.000
x2	0.510	0.416	0.662
T-Value	0.70	2.24	14.44
P-Value	0.501	0.052	0.000
x3	0.10		
T-Value	0.14		
P-Value	0.896		
x4	-0.14	-0.24	
T-Value	-0.20	-1.37	
P-Value	0.844	0.205	
S	2.45	2.31	2.41
R-Sq	98.24	98.23	97.87
R-Sq(adj)	97.36	97.64	97.44
Mallows C-p	5.0	3.0	2.7

Figure 10.9 Backward selection results from Minitab for the Hald cement data.

GAMBAR 10.9

Dengan model tiga variabel yang melibatkan (x_1, x_2, x_4). Dengan t statistik terkecil dalam model ini, t = -1,37 berhubungan dengan x_4. Karena |t| = 1,37 < $t_{OUT} = t_{\frac{0,20}{2},9}$ = 1,83, maka x_4 dikeluarkan dari model. Dan selanjutnya dengan melihat hasil fitting model dua variabel yang melibatkan (x_1, x_2). T statistik terkecil dalam model adalah 12,41 yang berhubungan dengan x_1 dan karena melebihi $t_{OUT} = t_{\frac{0,10}{2},10} = 1{,}81$ maka tidak ada lagi *regressor* yang bisa dikeluarkan oleh model. Sehingga, diperoleh model akhir.

$$\hat{y} = 52{,}5773 + 1{,}4683x_1 + 0{,}6623x_2.$$

Stepwise Regression **/ Regresi Bertahap** dimana asumsi ini dimulai modifikasi seleksi maju dimana pada setiap langkah, semua *regressor* yang dimasukkan kedalam model sebelumnya dinilai kembali melalui statistik F (atau t) parsial. Dengan syarat, jika jika statistik F parsial (atau t) untuk sebuah variabel kurang dari F_{OUT} (atau t_{OUT}) maka variabel tersebut dikeluarkan dari model. Dalam regresi bertahap membutuhkan dua nilai batas, yakni satu untuk memasukkan variabel dan satu lagi untuk mengeluarkannya.

EXAMPLE **/ CONTOH 10.4** ***STEPWISE REGRESSION – HALD CEMENT DATA*** **/ REGRESI BERTAHAP DATA SEMEN HALD**

Dengan melakukan penerapan regresi bertahap pada data semen hald yang diberikan dicontoh 10.1. Pada perhitungan ini diterapkan nilai $\alpha = 0{,}15$ untuk menambah atau menghapus *regressor*. Langkah ini dimulai dengan menambahkan x_4, dikarenakan statistik t melebihi $t_{IN} = t_{\frac{0,15}{2}}, 10 = 1{,}56$, maka x_4 dilakukan penghapusan. Namun, nilai t untuk

x_4 pada langkah berikutnya adalah -12,62, sehingga x_4 dipertahankan atau dilakukan penambahan.

Selanjutnya adalah proses menambahkan x_2 kedalam model. Kemudian dilakukan perbandingan antara statistik t untuk x_1 dan x_4 dibandingkan dengan $t_{OUT} = t_{\frac{0,15}{2}}, 9 = 1,57$. Karena nilai |t| $x_4 = 1,37 < t_{OUT} = 1,57$, maka x_4 dilakukan penghapusan. Dan karena langkah terakhir menunjukkan penghapusan pada x_4 maka kandadiat *regressor* yang tersisa adalah x_3 yang tidak dapat ditambahkan karena tidak memenuhi syarat. Sehingga, diperoleh model akhir.

$$\hat{y} = 52,5773 + 1,4683x_1 + 0,6623x_2.$$

Dimana persamaan model akhir tersebut sama dengan model akhir pada eliminasi mundur. Dan diperoleh hasil perhitungan Minitab, seperti dibawah ini.

Stepwise Regression: y versus x1, x2, x3, x4

Alpha-to-Enter: 0.15 Alpha-to-Remove: 0.15

Response is y on 4 predictors, with N = 13

Step	1	2	3	4
Constant	117.57	103.10	71.65	52.58
x4	-0.738	-0.614	-0.237	
T-Value	-4.77	-12.62	-1.37	
P-Value	0.001	0.000	0.205	
x1		1.44	1.45	1.47
T-Value		10.40	12.41	12.10
P-Value		0.000	0.000	0.000
x2			0.416	0.662
T-Value			2.24	14.44
P-Value			0.052	0.000
S	8.96	2.73	2.31	2.41
R-Sq	67.45	97.25	98.23	97.87
R-Sq(adj)	64.50	96.70	97.64	97.44
Mallows C-p	138.7	5.5	3.0	2.7

Figure 10.10 Stepwise selection results from Minitab for the Hald cement data.

GAMBAR 10.10

***General Comments On Stepwise-Type Procedures* / Komentar Umum Tentang Prosedur Tipe Bertahap**

Algortima regresi bertahap tidak menjamin identifikasi model regresi subset terbaik. Urutan masuk atau keluarnya *regressor* tidak selalu mencerminkan tingkat kepentingannya, proses ini juga dapat menghasilkan model akhir yang berbeda tergantung pada metode yang digunakan. Hal ini terlihat, pada data semen hald, dimana pada seleksi maju memilih x_4 sebagai regressor pertama, tetapi ketika x_2 ditambahkan, x_4 menjadi tidak berarti karena interkorelasi tinggi. Hal ini merupakan masalah umum dalam seleksi maju, di mana regressor yang dimasukkan tidak dapat dihilangkan pada langkah selanjutnya. Perlu diingat bahwa seleksi maju, eliminasi mundur, dan regresi bertahap

dapat menghasilkan pilihan model akhir yang berbeda karena interkorelasi antar *regressor* mempengaruhi urutan pemasukan dan pengeluaran.
Dengan ilustrasi perbedaan pada setiap prosedur menggunakan data semen hald, seperti berikut.

Pemilihan Ke Depan	x_1	x_2	x_4
Eliminasi Mundur	x_1	x_2	
Regresi Bertahap	x_1	x_2	

Berk [1978] mencatat bahwa seleksi maju cenderung cocok dengan semua regresi yang mungkin untuk ukuran subset kecil, tetapi tidak untuk yang besar, sementara eliminasi mundur cenderung cocok untuk ukuran subset besar, tetapi tidak untuk yang kecil. Oleh karena itu, beberapa pengguna merekomendasikan penerapan berbagai prosedur seleksi untuk memperoleh wawasan lebih mendalam tentang struktur data. Meskipun demikian, algoritma regresi bertahap diikuti dengan eliminasi mundur sering dianggap lebih stabil terhadap struktur korelatif dari regressor dibandingkan dengan seleksi maju.

Stopping Rules For Stepwise-Type Procedures / Aturan Penghentian Untuk Prosedur Bertahap

Penentuan nilai batas F_{IN} (atau t_{IN}) dan atau F_{OUT} (atau t_{OUT}) dalam prosedur tipe prosedur bertahap atau *stepwise* adalah aturan berhenti untuk algoritma tersebut. Beberapa program komputer memungkinkan pengguna untuk menentukan nilai ini langsung, sementara yang lain memerlukan pilihan tingkat kesalahan tipe 1α untuk menghasilkan nilai batas. Namun, karena nilai F parsial (atau t) yang dievaluasi pada setiap tahap adalah maksimum dari beberapa variabel F parsial (atau t) yang berkorelasi, menganggap α sebagai tingkat signifikansi atau tingkat kesalahan tipe 1 bisa menjadi keliru.

Penelitian oleh Draper, Guttman, Kanemasa, Pope, dan Webster belum membuat kemajuan signifikan dalam menentukan kondisi di mana tingkat signifikansi pada statistik t atau F menjadi bermakna, atau dalam mengembangkan distribusi persis dari statistik F atau t untuk masuk dan keluar dari model. Beberapa pengguna cenderung memilih nilai F_{IN} dan F_{OUT} (atau t yang setara) yang kecil agar bisa menginvestigasi *regressor* tambahan yang mungkin ditolak oleh nilai F yang lebih konservatif. Namun, strategi ekstrem seperti memilih nilai F_{IN} dan F_{OUT} sehingga semua *regressor* dimasukkan melalui seleksi maju atau dikeluarkan melalui seleksi mundur sehingga menghasilkan satu model subset untuk p = 2,3,..,k+1. Yang dapat dievaluasi dengan kriteria, seperti C_p atau MS_{Res} untuk menentukan model akhir.

Sebuah prosedur umum adalah menetapkan $F_{IN} = F_{OUT} = 4$, yang kira-kira sesuai dengan titik 5% teratas dari distribusi F. Alternatifnya, beberapa analis melakukan beberapa percobaan dengan nilai *cutoff* yang berbeda untuk mengamati efek pilihan kriteria pada subset yang dihasilkan. Studi-studi telah dilakukan untuk memberikan panduan praktis dalam pemilihan aturan berhenti. Bendel dan Affifi merekomendasikan α = 0,25 untuk seleksi maju, yang merupakan default dalam Minitab. Kennedy dan Bancroft juga

memberikan saran terkait pemilihan $\alpha = 0{,}25$ untuk seleksi maju dan $\alpha = 0{,}10$ untuk seleksi mundur. Pemilihan nilai cutoff seringkali tergantung pada preferensi pribadi dan terdapat kebebasan dalam penentuannya. Untuk diagram alir proses pembuatan model dapat direpresentasikan, seperti dibawah ini.

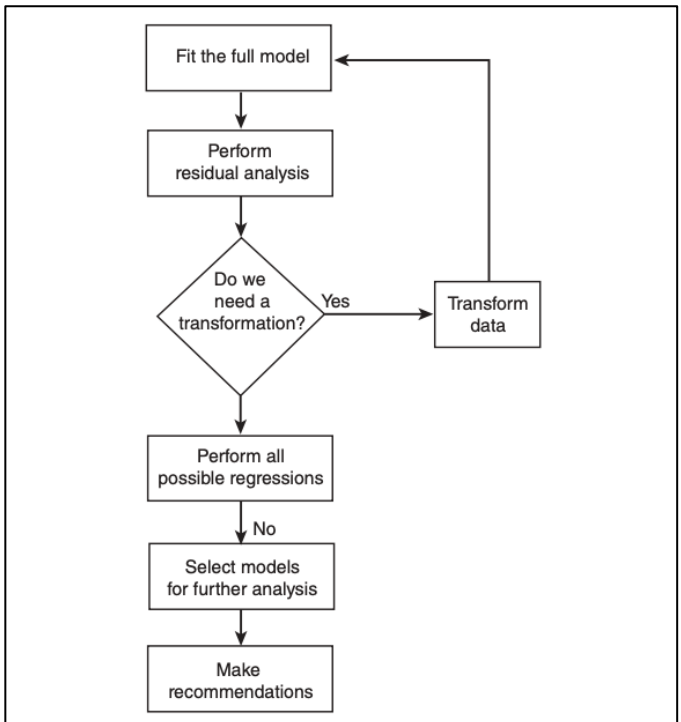


Figure 10.11 Flowchart of the model-building process.

GAMBAR 10.11 DIAGRAM ALIR PROSES PEMBUATAN MODEL

Diagram alir diatas merepresentasikan proses pembuatan model regresi linier.

- Prosesnya dimulai dengan memasukkan semua variabel prediktor ke dalam model. Variabel prediktor adalah variabel yang dapat digunakan untuk memprediksi variabel respon.
- Kemudian, prosedur pemilihan variabel diterapkan untuk menentukan variabel mana yang harus dimasukkan atau dihapus dari model. Prosedur pemilihan variabel yang umum digunakan adalah eliminasi ke belakang.
- Pada seleksi mundur, variabel prediktor dengan pengaruh terkecil terhadap variabel respon dihapus dari model. Prosedur ini diulangi sampai variabel prediktor yang tersisa tidak memiliki pengaruh signifikan terhadap variabel respon.
- Diagram alir ini menunjukkan langkah-langkah yang terlibat dalam proses pemilihan variabel untuk model regresi linier.

  Untuk penjelasan lebih rinci mengenai setiap langkah-langkah dalam pemilihan variabel untuk model regresi linier, dapat dijelaskan seperti berikut.
- Masukkan Semua Variabel Prediktor Ke Dalam Model

Pada langkah ini, semua variabel prediktor dimasukkan ke dalam model. Variabel prediktor dapat berupa variabel kuantitatif atau kategoris.

- Terapkan Prosedur Pemilihan Variabel
  Pada langkah ini, prosedur pemilihan variabel diterapkan untuk menentukan variabel mana yang harus dimasukkan atau dihapus dari model. Prosedur pemilihan variabel yang umum digunakan adalah eliminasi ke belakang.
- Hapus Variabel Prediktor Dengan Pengaruh Terkecil Terhadap Variabel Respon
  Pada langkah ini, variabel prediktor dengan pengaruh terkecil terhadap variabel respon dihapus dari model. Prosedur ini diulangi sampai variabel prediktor yang tersisa tidak memiliki pengaruh signifikan terhadap variabel respon. Langkah-langkah tersebut dapat diulangi hingga tidak ada lagi variabel prediktor yang dapat dihapus. Model akhir yang dihasilkan adalah model dengan variabel prediktor yang paling relevan untuk memprediksi variabel respon.

10.3 *STRATEGY FOR VARIABLE SELECTION AND MODEL BUILDING* / STRATEGI UNTUK PEMILIHAN VARIABEL DAN PEMBANGUNAN MODEL

Gambar 10.11 meringkas pendekatan dasar untuk pemilihan variabel dan oembuatan model. Langkah-langkah dasarnya adalah sebagai berikut:

1. Sesuaikan midel terbesar yang mungkin dengan data.
2. Lakukan analisis menyeluruh terhadap model.
3. Tentukan apakah transformasi respon atau beberapa regressor diperlukan.
4. Tentukan apakah semua kemungkinan regresi dapat dilakukan.
   - Jika semua kemungkinan regresi layak dilakukan, lakukan semua kemungkinan regresi dengan menggunakan kriteria seperti *Mallow's* C_p *adjusted* R^2, dan statistik PRESS untuk mengurutkan model subset terbaik.
   - Jika semua kemungkinan regresi tidak layak, gunakan teknik seleksi bertahap untuk menghasilkan model terbesar sehingga semua kemungkinan regresi dapat dilakukan.
5. Lakukan semua regresi yang mungkin seperti yang diuraikan di atas.
6. Bandingkan dan bedakan model terbaik yang direkomendasikan oleh masing-masing kriteria.
7. Lakukan analisis menyeluruh terhadap model-model "terbaik" (biasanya tiga sampai lima model).
8. Jelajahi kebutuhan untuk transformasi lebih lanjut.
9. Diskusikan dengan para ahli materi pelajaran tentang keuntungan dan kerugian relatif dari rangkaian model akhir.

   Alasan utama untuk menganalisis model lengkap adalah untuk mendapatkan gambaran tentang “gambaran besar”.

   Sangat penting bagi analisis untuk mengenali bahwa ada dua alasan dasar mengapa seseorang mungkin memerlukan transformasi respon :

- Analis menggunakan "skala" yang salah untuk tujuan tersebut. Contoh utama dari situasi ini adalah data jarak tempuh bensin. Kebanyakan orang lebih mudah menginterpretasikan respons sebagai "mil per-galon". Namun, data tersebut sebenarnya diukur sebagai "galon per-mil." Untuk banyak data teknik, skala yang tepat melibatkan transformasi log.
- Terdapat pencilan yang signifikan dalam data, terutama yang berkaitan dengan kecocokan dengan model penuh. Pencilan menunjukkan kegagalan model untuk menjelaskan beberapa respons.

Dalam beberapa kasus, respon itu sendiri yang menjadi masalah, misalnya, ketika respon tersebut salah diukur pada saat pengumpulan data. Pada kasus lain, model itu sendiri yang menciptakan pencilan. Dalam kasus-kasus ini, menghilangkan salah satu *regressor* yang tidak penting sebenarnya dapat mengatasi masalah.

Kami merekomendasikan penggunaan semua regresi yang memungkinkan untuk mengidentifikasi model subset setiap kali memungkinkan. Dengan daya komputasi saat ini, semua regresi yang memungkinkan biasanya dapat dilakukan untuk 20-30 *regressor* kandidat, tergantung pada ukuran total set data. Penting untuk diingat bahwa semua regresi yang memungkinkan menyarankan model terbaik secara murni berdasarkan kriteria apa pun yang analis pilih. Untungnya, ada beberapa kriteria baik yang tersedia, terutama C p Mallow, R yang disesuaikan, dan statistik PRESS. Secara umum, statistik PRESS cenderung merekomendasikan model yang lebih kecil daripada C p Mallow, yang pada gilirannya cenderung merekomendasikan model yang lebih kecil daripada R yang disesuaikan. Analis perlu merenungkan perbedaan dalam model dengan mempertimbangkan setiap kriteria yang digunakan. Semua regresi yang memungkinkan secara inheren mengarah pada rekomendasi beberapa model kandidat, yang lebih memungkinkan ahli subjek membawa pengetahuannya untuk membantu memecahkan masalah. Sayangnya, tidak semua paket perangkat lunak statistik mendukung pendekatan semua-regresi.

Metode langkah demi langkah (*stepwise*) cepat, mudah diimplementasikan, dan mudah ditemukan dalam banyak paket perangkat lunak. Sayangnya, metode ini tidak merekomendasikan model subset yang secara mutlak terbaik dengan mengacu pada kriteria standar apa pun. Selain itu, metode ini, menurut sifatnya, merekomendasikan satu persamaan akhir yang pengguna yang kurang berpengalaman mungkin salah menganggap sebagai optimal dalam beberapa hal.

Metode langkah demi langkah (*stepwise*) cepat, mudah diimplementasikan, dan mudah ditemukan dalam banyak paket perangkat lunak. Sayangnya, metode ini tidak merekomendasikan model subset yang secara mutlak terbaik dengan mengacu pada kriteria standar apa pun. Selain itu, metode ini, menurut sifatnya, merekomendasikan satu persamaan akhir yang pengguna yang kurang berpengalaman mungkin salah menganggap sebagai optimal dalam beberapa hal.

Strategi dua tahap, direkomendasikan menggunakan strategi dua tahap ketika jumlah *regressor* kandidat terlalu besar untuk menggunakan pendekatan semua regresi pada awal. Tahap pertama menggunakan metode langkah demi langkah untuk "memisahkan" regresor kandidat, mengeliminasi yang jelas tidak memiliki efek. Kemudian direkomendasikan menggunakan pendekatan semua regresi pada kumpulan *regressor* kandidat yang sudah direduksi. Ketika dihadapkan pada daftar regresor kandidat yang besar, biasanya bermanfaat untuk berinvestasi dalam pemikiran serius sebelum menggunakan komputer. Seringkali, kita dapat mengeliminasi beberapa *regressor* berdasarkan logika atau rasa teknik.

Penerapan yang benar dari pendekatan semua regresi seharusnya menghasilkan tiga hingga lima model kandidat akhir. Pada titik ini, sangat penting untuk melakukan analisis residu dan diagnostik menyeluruh untuk setiap model akhir ini. Saat membuat evaluasi akhir dari model yang dihasilkan, disarankan untuk bisa mengajukan pertanyaan, seperti mengenai pemeriksaan diagnostic yang biasa apakah bisa untuk mendapatkan model yang memuaskan, apakah terdapat persamaan yang tampak yang paling masuk akal, apakah *regressor* dalam model terbaik bisa mempertimbangkan masalah yang ada, model manakah yang cocok untuk digunakan dengan tujuan yang dimaksudkan, apakah koefisien regresi yang dihasilkan sudah sesuai, dan apakah masih terdapat masalah *multicollinearity.*

Jika empat pertanyaan ini dianggap serius dan jawabannya diterapkan dengan ketat, dalam beberapa kasus (mungkin banyak), tidak akan ada persamaan regresi akhir yang memuaskan. Misalnya, metode pemilihan variabel tidak menjamin memperbaiki semua masalah dengan *multicollinearity* dan pengaruh. Namun, terdapat situasi di mana *regressor* yang sangat terkait masih memberikan kontribusi signifikan ke model meskipun terkait. Ada titik data tertentu yang selalu tampak bermasalah. Analisis perlu mengevaluasi semua pertimbangan dalam membuat rekomendasi tentang model akhir. Secara jelas, penilaian dan pengalaman dalam lingkungan operasi yang dimaksudkan untuk model harus memandu analis saat membuat keputusan tentang model yang direkomendasikan.

Terakhir, beberapa model yang sesuai dengan data di mana mereka dikembangkan mungkin tidak memprediksi observasi baru dengan baik. Kami merekomendasikan agar analis menilai kemampuan prediktif model dengan mengamati kinerjanya pada data baru yang tidak digunakan untuk membangun model.