BAB I

PENDAHULUAN

Penerapan statistik dalam berbagai bidang, seringkali berdasar pada model-model statistik. Menurut Shao,J dan Tu,D (1995) salah satu model yang sering digunakan adalah model linier:

$$y_i = x_i' \beta + \varepsilon_i, \quad i = 1, \ldots, n. \qquad (1.1)$$

dengan y_i adalah variabel tak bebas atau respon ke-i , x_i merupakan p-vektor kolom variabel bebas atau prediktor yang berkaitan dengan y_i, β adalah p-vektor parameter model yang belum diketahui nilainya dan ε_i adalah galat yang diasumsikan bahwa ε_i independen dengan rataan 0 dan varian σ^2.

Parameter β dalam model tersebut dikenal sebagai parameter regresi. Dalam model regresi estimasi parameter dapat ditentukan dengan menggunakan metode kuadrat terkecil dan diperoleh nilai estimasi β sebagai berikut:

$$\hat{\beta} = (X'X)^{-1}X'y$$

dengan $X' = (x_1, x_2, \ldots, x_n)$ dan $y = (y_1, y_2, \ldots, y_n)'$.

Apabila estimasi parameter telah ditentukan, maka diperoleh estimasi model untuk respon y yang tergantung pada prediktor x, yang dapat digunakan untuk melakukan prediksi untuk nilai y yang akan datang berdasarkan prediktor x. Tetapi beberapa komponen dari x kemungkinan tidak menghasilkan prediksi yang akurat karena tidak berpengaruh secara signifikan terhadap respon y, maka perlu dilakukan

pemilihan model terbaik sedemikian hingga memiliki kemampuan prediksi yang lebih akurat.

Shao dan Tu (1995) menyebutkan bahwa terdapat beberapa prosedur pemilihan variabel dalam model regresi linier di antaranya adalah: leave-one-out Cross-Validation (CV-1), Metode Cp, dan Akaike Information Criterion (AIC). Permasalahan dalam memilih model terbaik terletak pada ukuran kemampuan prediksi yang terbaik (model yang kompak) dalam model linier tersebut, yaitu model yang memiliki rata-rata kuadrat galat terkecil / minimum.Selain dapat menentukan model terbaik, menurut Shao (1997) ketiga metode tersebut tercakup dalam satu kelompok berdasarkan ekuivalensinya.

Karena beberapa komponen dari β mungkin adalah sama dengan 0 maka model yang mempunyai kemampuan prediksi terbaik memiliki bentuk:

$$y = x'_{\alpha}\beta_{\alpha} + e \qquad (1.2)$$

dengan α adalah himpunan bagian dari $\{1, \ldots, p\}$.

Menurut Shao (1993) CV-1 adalah metode untuk pemilihan model-model regresi sesuai dengan ukuran kemampuan prediksi model tersebut hingga diperoleh satu model terbaik dengan melakukan estimasi parameter setelah menghilangkan satu data ke-i dan kemudian dipilih model yang memiliki rata-rata kuadrat galat minimum. Misal terdapat n data yang tersedia untuk memilih sebuah model dari sekelompok model yang tersedia. Data tersebut dibagi menjadi dua bagian, bagian pertama terdiri dari n_c data yang digunakan untuk pembentukan model (konstruksi model), dan

bagian kedua terdiri dari satu data yang digunakan untuk menaksir kemampuan prediksi dari model (validasi model). Untuk itu terdapat $\binom{n}{1}$ cara yang berbeda untuk membagi data yang tersedia. CV-1 memilih model dengan rata-rata ukuran kemampuan prediksi terbaik. Rata-rata kuadrat galat dapat dihitung dengan :

$$n^{-1}\left\| y_s - \hat{y}_{\alpha,s^c} \right\|^2 .$$

Dalam metode Cp, pemilihan variabel dilakukan dengan meminimalkan persamaan:

$$\Gamma_{n,\lambda}(\alpha) = \frac{S(\alpha)}{n} + \frac{\lambda \hat{\sigma}^2 p_\alpha}{n} \qquad (1.3)$$

dengan $\alpha \in A$, $S(\alpha) = \left\| y - x_\alpha' \hat{\beta}_\alpha \right\|$, $\hat{\sigma}^2$ adalah estimator dalam σ^2 dan $\{\lambda\}$ adalah barisan bilangan terurut yang bernilai lebih besar atau sama dengan 2.

Persamaan (1.3) tersebut merupakan persamaan GIC_λ. Apabila $\lambda \approx 2$ dikenal dengan metode Cp, untuk $\lambda \to \infty$ dikenal dengan metode GIC Rao dan Wu, dan $\lambda > 2$ merupakan metode FPE_λ. Dalam penulisan ini, pembahasan dibatasi hanya untuk nilai $\lambda \approx 2$.

Untuk metode AIC, dilakukan dengan meminimalkan persamaan

$$\tilde{\Gamma}_{n,\lambda}(\alpha) = \frac{S(\alpha)}{n}\left[1 + \frac{\lambda p_\alpha}{n - p_\alpha} \right]. \qquad (1.4)$$

Meminimalkan persamaan (1.4) dengan nilai $\lambda \approx 2$ dikenal dengan metode AIC dan apabila $\lambda = \log n$ merupakan metode BIC. Pembahasan selanjutnya dibatasi hanya untuk nilai $\lambda \approx 2$.

Model yang baik sesuai dengan persamaan (1.2), menurut Shao (1993) adalah model yang memiliki nilai rata-rata kuadrat galat terkecil dan jumlah variabel yang dilibatkan dalam model tersebut adalah yang sesedikit mungkin.

Untuk mendukung pernyataan-pernyataan di atas dilakukan simulasi terhadap "data konsumsi ban" yang diambil dari Draper, 1992, menggunakan program S-PLUS 2000.