Jurnal Ilmiah Matematika dan Pendidikan Matematika (JMP)
Vol. 10 No. 2, Desember 2018, hal. 43-48
ISSN (Cetak) : 2085-1456; ISSN (*Online*) : 2550-0422

PENENTUAN MODEL TERBAIK REGRESI RIDGE DAN TERAPANNYA

Sri Utami Zuliana

UIN Sunan Kalijaga
sri.zulianai@uin-suka.ac.id

ABSTRACT. *Ridge regression is one of penalized regression methods. Penalized regression methods are usually used for solving the problem of multicollinearity. The best model in ridge regression has been chosen by some previous techniques. In the techniques there is bias-variance trade-off. In this paper, Schall algorithm will be applied for choosing the best model. Schall algorithm is faster because it only needs a few iteratives to be convergence.*

Keywords*: penalized regression, ridge regression, the best model, Schall algorithm, the weights of penalty*

ABSTRAK. Regresi ridge merupakan salah satu metode regresi terpenalti. Metode regresi terpenalti biasanya digunakan untuk mengatasi multikolinear. Untuk menentukan model yang terbaik pada regresi ridge telah terdapat beberapa metode yang diperkenalkan. Penentuan model terbaik melibatkan pertimbangan antara bias atau variansi yang kecil. Dalam makalah ini akan diterapkan pemilihan model terbaik dengan menggunakan algoritma Schall. Algoritma Schall lebih cepat dalam memilih bobot penalti karena hanya memerlukan sedikit pengulangan untuk mencapai konvergen.

Kata Kunci: regresi terpenalti, regresi ridge, model terbaik, algoritma Schall, bobot penalty

1. PENDAHULUAN

Metode kuadrat terkecil berusaha meminimalkan error kuadrat ($\sum_i e_i^2$). Kecocokan model dapat sangat dipengaruhi oleh observasi yang memiliki jangkauan dan error yang besar. Bobot pada fungsi yang memberikan error yang besar dapat diminimalisir. Metode likelihood terpenalti adalah salah satu metode yang digunakan untuk mengurangi variansi pada data yang memiliki multikolinearitas.

Pemilihan model terbaik pada metode likelihood terpenalti melibatkan penawaran antara meminimalisir bias atau meminimalisir variansi. Ada beberapa metode yang sering diterapkan untuk pemilihan model terbaik yaitu Akaike Information Criterion (AIC) (Akaike, 1974), Bayesian Information Criterion (BIC) (Schwarz, 1978) dan Generalized Cross Validation (GCV) (Golub et al.,

1979). Regresi ridge adalah salah satu metode likelihood terpenalti (Hoerl dan Kennard, 1970). Pemilihan model terbaik dapat dilakukan dengan menggunakan algorithma Schall (Zuliana S.U. dan Perperoglou A., 2016). Pada makalah ini akan diterapkan pemilihan model terbaik regresi ridge pada data `prostate`.

2. DASAR TEORI

Metode kuadrat terkecil biasanya digunakan untuk mengestimasi koefisien-koefisien pada persamaan regresi. "Metode likelihood terpenalti adalah salah satu metode yang digunakan untuk mengurangi variansi pada data yang memiliki multikolinearitas."

Pada model linear tergeneralisasi, metode regularisasi likelihood terpenalti memodifikasi fungsi log-likelihood dengan menambahkan suatu suku penalty. Penambahan ini memperkecil variansi yang dimiliki oleh estimasi likelihood maksimum sehingga fungsinya menjadi lebih halus. Misal dimiliki fungsi log-likelihood $L(\beta)$ maka pada fungsi tersebut ditambahkan fungsi penghalus $s(\beta)$ menjadi

$$L^*(\beta) = L(\beta) - s(\beta).$$

Salah satu metode likelihood terpenalti menggunakan fungsi penghalus norm L_q

$$s(\beta) = \lambda \sum_{j=1}^{p} \left|\beta_j\right|^q,$$

di mana $q \geq 0$ dan $\lambda \geq 0$. Konstanta λ disebut parameter penghalus karena derajat kehalusan bergantung pada konstanta ini.

3. METODE PEMILIHAN MODEL TERBAIK REGRESI RIDGE

Regresi ridge diperkenalkan oleh Hoerl dan Kennard pada tahun 1970. Regresi ridge didefinisikan dengan:

$$s(\beta) = \frac{1}{2} \lambda \sum_{j=1}^{p} \left|\beta_j\right|^2$$

Dapat dilihat bahwa regresi ridge adalah metode likelihood terpenalti menggunakan fungsi penghalus norm L_2.

Pemilihan konstanta λ merupakan tawar menawar antara bias dan variansi. Semakin besar λ maka estimasi koefisien $\hat{\beta}_j$ semakin mendekati nol sehingga variansi semakin kecil tetapi biasnya semakin besar. Ada beberapa metode yang sering digunakan seperti AIC (Akaike, 1974), BIC (Schwarz, 1978) dan GCV (Golub et al., 1979).

Algoritma Schall diperkenalkan pertama kali untuk mengestimasi variansi dari efek acak (Schall, 1991). Zuliana dan Perperoglou menawarkan algoritma Schall untuk mengestimasi bobot penalti yang optimal (Zuliana dan Perperoglou, 2016). Algoritma tersebut memiliki tahap sebagai berikut:

1. Algoritma dimulai dengan sembarang $\hat{\lambda}$ dan mengestimasi koefisien $\hat{\beta}$:

$$\hat{\beta} = (X^T \tilde{W} X + \hat{\lambda} I)^{-1} X^T \tilde{W} \tilde{z}$$

2. Dari estimasi koefisien $\hat{\beta}$ yang diperoleh dapat digunakan untuk mengestimasi variansi:

$$\hat{\sigma} = \frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{n - ED}$$

dan

$$\hat{\tau} = \frac{\beta^T \beta}{ED}$$

sehingga estimasi $\hat{\lambda}$ yang baru diperoleh dengan:

$$\hat{\lambda} = \frac{ED}{\beta^T \beta}$$

3. Pengiterasian terus dilakukan sampai didapatkan $\hat{\lambda}$ yang konvergen.

Penjelasan yang lebih rinci tentang algoritma ini dapat dibaca di Zuliana dan Perperoglou (2016) dan Zuliana (2017).

4. PENERAPAN PEMILIHAN MODEL TERBAIK DENGAN MENGGUNAKAN ALGORITMA SCHALL

Dimiliki data prostate yang digunakan untuk mencari hubungan antara level *prostate-spesific antigen* (`lpsa`) dengan beberapa indikator kesehatan yang diukur dari para laki-laki yang menerima perlakuan *radical prostatectomy*.

Indikator-indikator yang diukur adalah *log cancer volume* (`lcavol`), *log prostate weight* (`lweight`), usia dalam tahun (`age`), *log of the amount of benign prostatic hyperplasia* (`lbph`), *seminal vesicle invasion* (`svi`), *log of capsular penetration* (`lcp`), *Gleason score* (`gleason`), dan *percent of Gleason score 4 or 5* (`pgg45`). Variabel-variabel ini memiliki korelasi yang cukup tinggi, bisa dilihat di Tabel 1.

Apabila dibentuk regresi linear didapatkan model yang memiliki P-value sangat signifikan tetapi beberapa variabelnya memiliki P-value yang tidak signifikan. Ini dapat dilihat pada Tabel 2.

Kemudian dibentuk regresi ridge dengan menggunakan algoritma Schall. Dilakukan pemilihan model terbaik dengan menggunakan AIC, BIC, GCV dan algoritma Schall. Untuk bobot penalti optimal AIC, BIC dan GCV didapatkan hasil sebagai berikut 5,75 19,95 dan 6,31. Sedangkan dengan algoritma Schall, didapatkan bobot penalti optimal adalah 19,98 sehingga didapatkan model:

$$\hat{y} = 0{,}4360\,lcavol + 0{,}1783\,lweight - 0{,}0658\,age + 0{,}1021\,lbph$$
$$+ 0{,}2193\,svi + 0{,}0256\,lcp + 0{,}0455\,gleason + 0{,}0692\,pgg45 .$$

Sesatan kuadrat rata-rata yang didapatkan dari AIC, BIC, dan GCV hampir sama yaitu 33,44; 34,98; dan 33,49 . Sedangkan sesatan rata-rata yang dimiliki oleh algoritma Schall adalah 34,98. Kelebihan dari algoritma Schall adalah pemilihan bobot penalty awal $\hat{\lambda}_0$ dapat dilakukan oleh sembarang bilangan dan iterasi yang dilakukan hingga mendapatkan $\hat{\lambda}$ yang konvergen biasanya tidak terlalu banyak.

Tabel 1. Korelasi variabel-variabel pada data prostate

Tingkat Korelasi	Lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45
Lcavol	1,00	0,19	0,22	0,03	0,54	0,68	0,43	0,43
Lweight	0,19	1,00	0,31	0,43	0,11	0,10	0,00	0,05
Age	0,22	0,31	1,00	0,35	0,12	0,13	0,27	0,28
Lbph	0,03	0,43	0,35	1,00	-0.09	-0.01	0.08	0.08
Svi	0,54	0,11	0,12	-0,09	1,00	0,67	0,32	0,46
Lcp	0,68	0,10	0,13	-0,01	0,67	1,00	0,51	0,63
Gleason	0,43	0,00	0,27	0,08	0,32	0,51	1,00	0,75
pgg45	0,43	0,05	0,28	0,08	0,46	0,63	0,75	1,00

Tabel 2. Koefisien estimasi hasil dari regresi linear pada data prostate

	Estimate	P-value
lcavol	0,5994	0,0000
lweight	0,1955	0,0086
Age	-0,1267	0,0806
Lbph	0,1346	0,0688
Svi	0,2748	0,0022
Lcp	-0,1278	0,2470
gleason	0,0282	0,7738
pgg45	0,1106	0,3061

5. KESIMPULAN DAN SARAN

Algoritma Schall telah diperkenalkan sebagai salah satu metode pemilihan model terbaik regresi ridge (Zuliana dan Perperoglou, 2016). Dalam tulisan ini, algoritma tersebut diterapkan pada data prostate untuk menunjukkan bahwa algoritma ini memiliki kinerja yang hampir sama dengan metode pemilihan model terbaik yang lainnya.

Algoritma Schall sudah dipakai pada regresi spline (Rigby dan Stasinopolous, 2014). Ide yang sama juga disampaikan pada model PRIDE (Perperoglou dan Eilers, 2010). Algoritma Schall sudah diterapkan pada metode regresi terpenalti lainnya seperti pada model linear tergeneralisir, model aditif tergeneralisir, dan juga regresi lasso (Zuliana, 2017). Algoritma Schall juga sudah diterapkan untuk mencari fungsi penghalus fungsi dua variabel dengan penghalus Whittaker (Zuliana dan Perperoglou, 2017).

DAFTAR PUSTAKA

Agresti, A., *Foundations of Linear and Generalized Linear Models*, John Wiley & Sons, Inc., 2015.

Akaike, H., *A new look at the statistical model identification*, IEEE transactions on automatic control, **19**(6) (1974), 716-723.

Golub, G. H., Heath, M., & Wahba, G., *Generalized cross-validation as a method for choosing a good ridge parameter*, Technometrics, **21**(2) (1979), 215-223.

Hoerl, A. E., & Kennard, R. W., *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics, **12**(1) (1970), 55-67.

Schall, R., Estimation in Generalized Linear Models with Random Effects, Biometrika, **78**(4) (1991), 719-727.

Schwarz, G., *Estimating the dimension of a model*, The annals of statistics, **6**(2) (1978), 461-464.

Stamey, T.A., et al., *Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate: II. radical prostatectomy treated patients*, Journal of Urology, **141**(5) (1989), 1076–1083.

Zuliana, S. U., *Penalized Regression Methods with Application to Generalized Linear Models, Generalized Additive Models, and Smoothing*, Doctoral dissertation, University of Essex, 2017.

Zuliana S.U. dan Perperoglou A., *The Weight of Penalty Optimization for Ridge Regression*, dalam Wilhelm A., Kestler H. (eds), *Analysis of Large and Complex Data*, Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Cham, 2016, 231-239.

Zuliana, S. U., & Perperoglou, A., *Two dimensional smoothing via an optimised Whittaker smoother*, Big Data Analytics, **2**(1) (2017), 6.