Statistika, Vol. 6 No. 2, 35 – 39
Nopember 2006

Pengkelasan dengan Skor Propensitas

MARZUKI DAN FAKHRURRAZI

Jurusan Matematika FMIPA Unsyiah
Jl. Syech Abdul Rauf No. 3 Darussalam Banda Aceh 23111

ABSTRAK

Perbandingan dua populasi dengan latar belakang objek (kovariat-kovariat) yang bervariasi akan menghasilkan simpulan yang bias. Skor propensitas dapat digunakan sebagai salah satu solusinya. Pengkelasan objek dengan skor propensitas menghasilkan rata-rata persentase ketepatan pengkelasan dari 10 perulangan adalah 86.9% untuk satu kovariat dan bila dua kovariat maka rata-rata persentase ketepatan pengkelasan terhadap objek dari 10 perulangan adalah 88.4%.

1. Pendahuluan

Membandingkan suatu parameter dua kelompok akan sulit jika objek-objek dalam kedua kelompok tersebut berbeda dalam hal latar belakangnya. Hal ini disebabkan perbandingan tersebut mengasumsikan tidak ada pengaruh peubah lain (kovariat) atau dengan kata lain setiap objek yang diamati mempunyai latar belakang yang sama kecuali peubah yang diamati. Salah satu cara yang digunakan untuk mengendalikan perbedaan itu adalah dengan membagi latar belakang objek-objek tersebut menjadi kelas-kelas berdasarkan karakteristik yang diamati. Salah satu karakteristik itu adalah skor propensitas.

Penelitian ini bertujuan untuk mengurai suatu prosedur pengkelasan objek berdasarkan skor propensitas dan mengevaluasi metode pengkelasan skor propensitas menggunakan data simulasi.

Kajian tentang pengkelasan dengan skor propensitas ini dapat dipakai sebagai langkah awal untuk penelitian tentang perbandingan dua populasi supaya variasi yang disebabkan oleh latar belakang dari objek penelitian dapat direduksi.

Data yang dibangkitkan data yang menyebar normal. Pengkelasan ditentukan maksimal untuk tiga kelas dan kovariat yang ditetapkan sebanyak satu dan dua kovariat.

2. Tinjauan Pustaka

Regresi Logistik

Model regresi logistik adalah salah satu model regresi yang digunakan untuk melihat hubungan antara satu peubah respons dan satu atau beberapa peubah prediktor. Peubah responsnya berbentuk data biner sedangkan peubah-peubah prediktornya mungkin peubah kategorik atau peubah kontinu.

Misal data sampel dalam bentuk $\{y_i, x_{i1}, x_{i2}, \ldots, x_{im}; i = 1, 2, \ldots, n\}$. Data peubah y_i diasumsikan berasal dari data populasi yang berdistribusi Bernouli (p_i), sedangkan $x_{i1}, x_{i2}, \ldots, x_{im}$ adalah data sampel pengamatan ke-i untuk peubah prediktor $X_1, \ldots, X_m$ (m menyatakan banyaknya peubah prediktor). Bentuk dasar dari taksiran model regresi logistik adalah:

$$\text{logit}(\hat{p}_i) = \ln\left[\frac{\hat{p}_i}{1-\hat{p}_i}\right] = \hat{\beta}_0 + x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2 + \ldots + x_{im}\hat{\beta}_m; \quad i = 1,2,\ldots,n$$

sedangkan

$$\hat{p}_i = \frac{e^{\hat{\beta}_0 + x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2 + \ldots + x_{im}\hat{\beta}_m}}{1 + e^{\hat{\beta}_0 + x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2 + \ldots + x_{im}\hat{\beta}_m}}$$

dengan $\hat{p}_i$ menyatakan taksiran peluang sukses bagi pengamatan ke-*i*.

Taksiran Parameter Model Regresi Logistik

Fungsi kemungkinan dari parameter-parameter model regresi logistik untuk data sampel $\{y_i, x_{i1}, x_{i2}, ..., x_{im}\ ;\ i = 1, 2, ..., n\}$ adalah sebagai berikut

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{1-y_i}\ ;\ \text{di mana}\ \beta^T = (\beta_o, \beta_1, \beta_2, ..., \beta_m)$$

Bentuk logaritma natural fungsi kemungkinannya adalah:

$$\ln L(\beta) = \sum_{i=1}^{n} \left[y_i \ln p_i + (1-y_i) \ln(1-p_i) \right]$$

$$\ln L(\beta) = \sum_{i=1}^{n} \left[y_i \ln p_i - y_i \ln(1-p_i) + \ln(1-p_i) \right]$$

$$\ln L(\beta) = \sum_{i=1}^{n} \left[y_i \ln\left(\frac{p_i}{1-p_i} \right) + \ln(1-p_i) \right]$$

dengan:

$$p_i = \frac{e^{\beta_0 + \sum_{j=1}^{m} \beta_{ij} x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{m} \beta_{ij} x_{ij}}}$$

$$1 - p_i = \frac{1}{1 + e^{\beta_0 + \sum_{j=1}^{m} \beta_{ij} x_{ij}}}$$

$$\frac{p_i}{1-p_i} = \frac{\dfrac{e^{\beta_0 + \sum_{j=1}^{m} \beta_{ij} x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{m} \beta_{ij} x_{ij}}}}{\dfrac{1}{1 + e^{\beta_0 + \sum_{j=1}^{m} \beta_{ij} x_{ij}}}} = e^{\beta_0 + \sum_{j=1}^{m} \beta_{ij} x_{ij}}$$

$$\ln L(\beta) = \sum_{i=1}^{n} \left[(y_i \beta_0 + y_i \sum_{j=1}^{m} \beta_{ij} x_{ij}) - \ln\left(1 + e^{\beta_0 + \sum_{j=1}^{m} \beta_{ij} x_{ij}} \right) \right]$$

Taksiran parameter model regresi logistik $\hat{\beta}$ adalah nilai β yang memaksimumkan $\ln L(\beta)$. Nilai β yang membuat $\ln L(\beta)$ di atas maksimum tidak dapat diperoleh secara analitik. Hosmer dan Lemeshow (1989) menyatakan ini disebabkan pada saat turunan pertama dari $\ln L(\beta)$ = 0, yaitu

$$\sum_{i=1}^{n} (y_i - p_i) = 0 \ \text{dan}\ \sum_{i=1}^{n} x_i (y_i - p_i) = 0\ ,$$

tidak diperoleh solusi untuk β. Salah satu metode untuk memperoleh nilai taksiran parameter β di atas adalah menggunakan metode iterasi Newton-Raphson (Supardi, 2003). Nilai $\hat{\beta}$ yang diperoleh disebut penaksir kemungkinan maksimum dari parameter β.

Skor Propensitas

Gagasan awal dari metode ini adalah menggantikan koleksi dari kovariat-kovariat yang membaur dalam studi penelitian dengan satu fungsi dari kovariat-kovariat ini yang disebut skor propensitas (Rubin 1997). Jika skor propensitas ini digunakan maka akan sama halnya dengan melibatkan hanya satu kovariat. Skor propensitas diperoleh dengan memprediksi keanggotaan kelompok dari kovariat tersebut di mana setiap subjek pengamatan diduga satu skor propensitas. Skor ini merupakan peringkasan kovariat-kovariat menjadi satu yang digunakan untuk pengelompokan.

Metode skor propensitas diberlakukan untuk dua kelompok secara serentak. Setiap subjek individu yang menunjukkan indikator kelompok diberi notasi Z dengan Z=1 untuk kelompok ke-1 dan Z = 0 untuk kelompok ke-2, sedangkan vektor kovariat yang terdapat dalam kelompok diberi notasi **X**.

Nilai skor propensitas didefinisikan oleh Tu dan Zhou (2003) sebagai

$e(\mathbf{X}) = P(Z=1 \mid \mathbf{X})$

dan Sianesi (2001) menulis $0< e(\mathbf{X}) = P(Z=1 \mid \mathbf{X}=\mathbf{x}) <1$ untuk setiap $\mathbf{x} \in \mathbf{X}$ yang tidak lain adalah nilai peluang bersyarat suatu kelompok berdasarkan kovariat-kovariat yang diamati. Nilai ini dapat digunakan untuk mengendalikan bias karena ketidakseimbangan kovariat yang diamati. Parsons (2001) menyatakan klasifikasi skor propensitas (KSP) merupakan salah satu cara untuk mengurangi bias dari suatu kelompok yang melibatkan kovariat.

Pendugaan skor propensitas $e(\mathbf{X})$ dilakukan melalui model logistik

$$\ln\left(\frac{P(Z=1 \mid \mathbf{X}=\mathbf{x})}{1-P(Z=1 \mid \mathbf{X}=\mathbf{x})}\right) = \mathbf{x}^t\beta$$

sehingga penduga skor propensitas $e(\mathbf{X})$ adalah

$$\hat{e}(\mathbf{x}) = \frac{\exp(\mathbf{x}^t\hat{\beta})}{1+\exp(\mathbf{x}^t\hat{\beta})}$$

Semua subjek distratifikasi menjadi K kelas dengan menggunakan penduga skor propensitas sehingga setiap subjek dalam setiap kelas mempunyai nilai penduga skor propensitas yang hampir sama. Misalkan n_{1k} dan $Y_{1k1}, \ldots, Y_{1kn_{1k}}$ adalah jumlah subjek dan respon dalam kelompok ke-1 dalam kelas ke-k. n_{0k} dan $Y_{0k1}, \ldots, Y_{0kn_{ok}}$ adalah jumlah subjek dan respon dalam kelompok ke-2 dalam kelas ke-k; dengan $k = 1, \ldots, K$.

Jumlah total subjek dalam kelompok ke-1 dan kelompok ke-2 adalah

$$n_1 = \sum_{k=1}^{K} n_{1k}$$

dan

$$n_0 = \sum_{k=1}^{K} n_{0k}$$

Rosenbaum dan Rubin (1984) menyatakan bahwa lima kelas berdasarkan skor propensitas dapat menghilangkan bias lebih dari 90% karena kovariat yang tidak seimbang.

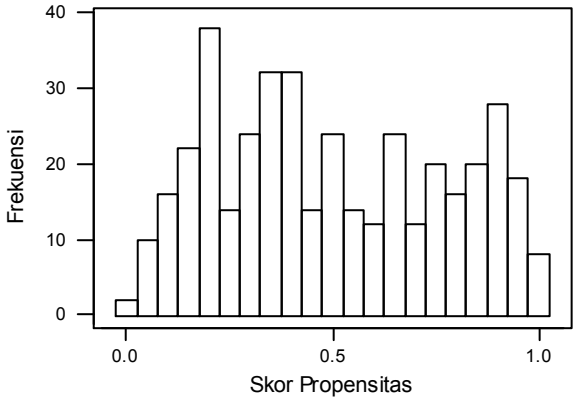
3. Hasil Data Simulasi

Kasus I ditetapkan satu kovariat X_1 untuk dua kelas yaitu kovariat untuk kelas satu, X_{11}~N(0, 1) dan kovariat untuk kelas dua, X_{12}~N(1, 1) dengan menggunakan program bangkitan data pada perangkat lunak *MATLAB*, menghasilkan 100 data sebagai objek. Hasil ini juga digunakan sekaligus guna mencari skor propensitas untuk setiap objek tersebut. Skor propensitas ini diurutkan dari kecil ke besar kemudian dibagi menjadi dua kelas.

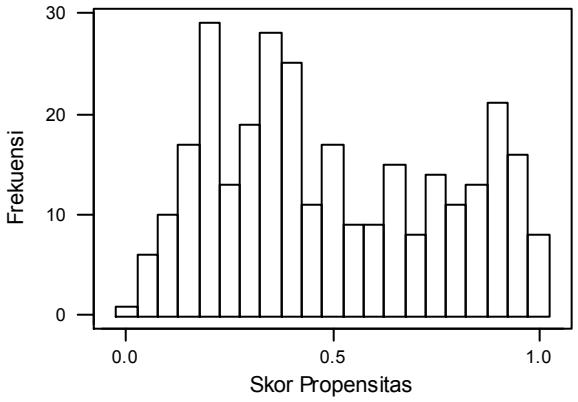
Kasus II, ditentukan dua kelas dan dua kovariat yaitu X_{11}~N(0, 0.5) dan X_{21}~N(0.5, 0.5) serta X_{12}~N(1, 1.5) dan X_{22}~N(1.5, 1.5). Penggunaan program bangkitan data, menghasilkan 100 data sebagai objek. Hasil ini juga digunakan sekaligus guna mencari skor propensitas untuk setiap objek tersebut. Skor propensitas ini diurutkan dari kecil ke besar kemudian dibagi menjadi dua kelas. Langkah-langkah ini diulang sebanyak 10 kali.

4. Penyebaran Skor Propensitas

Skor propensitas menyebar hampir sama antara objek-objek yang mempunyai satu kovariat dan objek-objek yang mempunyai dua kovariat. Ini dapat dilihat pada Gambar 1 dan Gambar 2 yang menyajikan frekuensi dari skor propensitas yang diwakili masing-masing untuk perulangan pertama. Dengan kata lain, Kasus I dan II mempunyai sebaran kovariat yang hampir sama.



Gambar 1. Sebaran skor propensitas untuk Kasus I ulangan pertama



Gambar 2. Sebaran skor propensitas untuk Kasus II ulangan pertama

5. Persentase Ketepatan Pengkelasan

Kasus I

Pengkelasan yang tepat untuk ulangan pertama sebanyak 86% dari 100 objek yang ditetapkan. Dengan kata lain, dari 100 objek itu hanya pada 14 objek terjadi kesalahan pengkelasan. Artinya 14 objek itu (objek ke-9, 17, 20, 23, 29, 30, 39, 41, 43, 44, 46, 47, 48, dan objek ke-50) ditempatkan pada kelas satu padahal kelas dua. Secara otomatis, karena pembagian menjadi dua kelas maka ada 14 objek lain ditempatkan pada kelas dua yang seharusnya pada kelas satu. Demikian pula untuk perulangan kedua sampai ke sepuluh maka kesalahanan pengkelasan masing-masing adalah 11%, 14%, 8%, 13%, 15%, 12%, 15%, 15%, dan 14%. Sehingga rata-rata persentase ketepatan pengkelasan terhadap objek dari 10 perulangan adalah 86.9%. Persentase kesalahan adalah 100% dikurangi dengan persentase ketepatan. Rata-rata persentase kesalahan pengkelasan terhadap objek dari 10 perulangan adalah 13.1%. Persentase ketepatan pengkelasan ini disajikan pada Tabel 1.

Tabel 1. Persentase ketepatan pengkelasan untuk kasus I

Ulangan	1	2	3	4	5	6	7	8	9	10
Persentase	86	89	86	92	87	85	88	85	85	86

Kasus II

Apabila menggunakan dua kovariat maka rata-rata persentase ketepatan pengkelasan terhadap objek dari 10 perulangan adalah 88.4%. Persentase ketepatan pengkelasan ini disajikan pada Tabel 2. Ini berarti hanya sekitar 11.6% rata-rata kesalahan pengkelasan. Penjelasan terhadap masing-masing ulangan itu dapat dianalogikan seperti pada Kasus I di atas.

Tabel 2. Persentase ketepatan pengkelasan untuk kasus II

Ulangan	1	2	3	4	5	6	7	8	9	10
Persentase	87	87	88	88	87	90	88	89	89	91

6. Penutup

Skor propensitas untuk kasus I yaitu satu kovariat X_1 untuk dua kelas dengan kovariat untuk kelas satu, X_{11}~N(0, 1) dan kovariat untuk kelas dua, X_{12}~N(1, 1) dan skor propensitas untuk kasus II yang ditentukan dua kelas dan dua kovariat yaitu X_{11}~N(0, 0.5) dan X_{21}~N(0.5, 0.5) serta X_{12}~N(1, 1.5) dan X_{22}~N(1.5, 1.5) memiliki sebaran yang hampir sama.

Pengkelasan untuk dua kelas dengan satu kovariat menghasilkan rata-rata persentase ketepatan pengkelasan terhadap objek dari 10 perulangan adalah 86.9%. Apabila menggunakan dua kovariat maka rata-rata persentase ketepatan pengkelasan terhadap objek dari 10 perulangan adalah 88.4%.

7. Daftar Pustaka

[1] Hosmer DW, Lemeshow S, 1989, *Applied Logistic Regression*. John Wiley & Sons, New York

[2] Myers RH, 1990, *Classical and Modern Regression with Applications*. PWS-KENT Publishing Company, Boston

[3] Parsons LS, 2001, Reducing Bias in a Propensity Score Matched-Pair Sample Using Greedy Matching Techniques. *Ovation Research Group* paper 214-26

[4] Rosenbaum PR, Rubin DB, 1983, Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 70:41-55

[5] Rosenbaum PR, Rubin DB, 1984, Reducing Bias in Observational Studies Using Subclassification on the Propensity Score, *Journal of the American Statistical Association* 79: 318-328

[6] Rubin DB, 1997, Estimation from Nonrandomized Treatment Comparisons Using Subclassification on Propensity Scores, *Nonrandomized Comparative Clinical Studies* pp. 757-763

[7] Tu W, Zhou XH, 2003, A Bootstrap Confidence Interval Procedure for the Treatment Effect Using Propensity Score Subclassification, *UW Biostatistics Working Paper Series* paper 200