Model Quasi-likelihood untuk Mengatasi Masalah Overdispersi pada Data yang Berdistribusi Multinomial

Uli Silma*, Nusar Hajarisman

Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung, Indonesia.

*ulisilma1@gmail.com, nusarhajarisman@yahoo.com

Abstract. Discrete data is data in the form of numbers (numbers) obtained by counting. As stated by McCullagh and Nelder (1989), overdispersion problems will often be encountered in discrete data analysis, both response variables in the form of binary (dichotomous), counts, and structures of more than two categories (polychotomous) such as in this multinomial distributed model.The number of events with more than two categories can be expressed by following a multinomial distribution. Several basic assumptions must be met when applying a multinomial distributed model to a particular data set, one of which is that the response variable is an independent random variable, and the probability of success of an event is constant. However, in practice, it is not uncommon for assumptions to be violated, in which random variables are independent. The independence between random variables is interpreted as a correlation between the response variables, this is an indication that there is a problem called overdispersion. Data can be said to have overdispersion problems when the Pearson deviance or chi-squared value is more than 1 (McCullagh & Nelder, 1989). The multinomial distribution is one method that can be used to overcome the problem of overdispersion in data that follows a normal distribution. The quasi-likelihood model discussed in this thesis is one model that can be used to overcome the problem of overdispersion in data that follows a multinomial distribution. The data used by the author for the application of the quasi-likelihood model is data on the satisfaction level of PDAM users in Antapani Wetan village in October 2019.

Keywords: *Polychotomous data, Multinomial Distribution, Overdispersion, Quasi-likelihood Model..*

Abstrak. Data diskrit adalah data yang berbentuk angka (bilangan) yang diperoleh dengan cara membilang. Sebagaimana yang diungkapkan oleh McCullagh dan Nelder (1989), masalah overdispersi akan sering dijumpai dalam analisis data diskrit, baik variabel respons yang berbentuk biner (dikotomus), cacahan, maupun struktur lebih dari dua kategori (polikotomus) seperti dalam model yang berdistribusi multinomial ini. Banyaknya kejadian dengan lebih dari dua kategorik dapat dinyatakan dengan mengikuti distribusi multinomial. Ada beberapa asumsi dasar yang harus dipenuhi pada saat mengaplikasikan model yang berdistribusi multinomial pada gugus data tertentu, salah satunya adalah bahwa variabel respon merupakan variabel acak yang saling bebas, dan peluang sukses dari suatu kejadian adalah konstan. Namun pada praktiknya tak jarang terjadi pelanggaran asumsi, di mana terdapat ketidakbebasan antar variabel acak. Ketidakbebasan antar variabel acak dimaknai sebagai adanya korelasi diantara variabel respon, hal tersebut merupakan suatu bukti indikasi bahwa terdapat masalah yang disebut overdispersi. Data dapat dinyatakan mengalami masalah overdispersi ketika nilai devians atau chi kuadrat Pearson yang lebih dari 1 (McCullagh & Nelder, 1989). Distribusi multinomial merupakan salah satu metode yang dapat digunakan untuk mengatasi masalah overdispersi pada data yang mengikuti distribusi normal. Model quasi-likelihood yang dibahas dalam skripsi ini merupakan salah satu model yang dapat digunakan untuk mengatasi masalah overdispersi pada data yang mengikuti distribusi multinomial. Data yang digunakan penulis untuk penerapan model quasi-likelihood adalah data tingkat kepuasan pengguna PDAM di kelurahan Antapani Wetan Bulan Oktober Tahun 2019.

Kata Kunci: *Data polikotomus, Distribusi Multinomial, Overdisperdi, Model Quasi-likelihood.*

Corresponding Author
Email: nusarhajarisman@yahoo.com

A. Pendahuluan

Munculnya masalah overdispersi tidak hanya terjadi dalam pengamatan data biner saja, tetapi juga muncul pada data diskrit lainnya dimana variabel acaknya mengikuti distribusi Poisson atau pun multinomial. Sedangkan masalah overdispersi dalam data diskrit itu sendiri dapat dijelaskan oleh dua hal, yaitu: adanya keragaman dalam peluang respon dan adanya korelasi antar peubah respon. Kedua kejadian tersebut merupakan kejadian yang bolak-balik, artinya apabila terdapat keragaman dalam peluang respon, maka terdapat korelasi antar peubah respon. Begitu juga sebaliknya, jika terdapat korelasi antara peubah respon, maka terdapat keragaman dalam peluang respon. McCullagh dan Nelder (1989) menyatakan bahwa kedua kejadian tersebut dapat terjadi karena adanya pengelompokan (clustering) dalam populasi. Rumah tangga, keluarga, litter, lingkungan, dan lain-lain, secara alami dapat membentuk sendiri kelompok-kelompoknya. Sedangkan Collet (1990) menyebutkan bahwa kejadian-kejadian itu muncul karena sejumlah unit percobaan diamati beberapa kali pada kondisi yang sama, sehingga akan diperoleh suatu peluang respon yang berbeda dari satu percobaan ke percobaan yang lainnya.

Konsekuensi dari adanya masalah overdispersi dalam data diskrit yang disebabkan oleh adanya keragaman dalam peluang respon serta adanya korelasi antara peubah respon adalah pada nilai penduga ragamnya. Apabila penduga ragam ini digunakan untuk menghitung selang kepercayaan dan untuk mengerjakan pengujian hipotesis statistik, maka akan diperoleh rataan yang terlalu kecil. Hal ini akan berakibat pada selang kepercayaan yang terlalu pendek serta pada pengujian hipotesis akan selalu menolak hipotesis nol. Dengan kata lain, dalam melakukan analisis untuk kasus seperti ini, maka hal ini akan memperbesar salah satu jenis, yang artinya peluang untuk menolak hipotesis nol yang seharusnya diterima menjadi semakin besar. Berdasarkan hal tersebut, maka perlu dicari suatu metode untuk mendapatkan solusi statistika yang tepat dalam menentukan hubungan fungsional antara respon dengan sejumlah peubah penjelas, dimana peubah responnya berkorelasi.

Ada beberapa macam model yang diusulkan oleh sejumlah penulis untuk menangani masalah overdispersi dalam data multinomial ini, diantaranya yaitu: model berbasis quasi-likelihood yang diperkenalkan oleh Wedderburn (1974), dan kemudian dikembangkan oleh McCullagh, dan Nelder (1989), serta pendekatan metode generalized estimating equations (GEE) yang dikembangkan Liang and Zeger (1986). Berbagai pendekatan lainnya untuk memodelkan data multinomial yang mempunyai masalah overdispersi ini adalah model Dirichlet-multinomial yang dibahasa oleh Mosimann (1962), generalized linear mixed models Wolfinger and O'connell (1993), serta finite-mixture model yang dikembangkan oleh Morel and Nagaraj (1993) Dalam skripsi ini akan dibahas mengenai model quasi-likelihood untuk menangani masalah overdispersi dalam data yang berdistribusi multinomial. Model ini akan diaplikasikan untuk melihat tingkat kepuasan konsumen pengguna layanan Perusahaan Daerah Air Minum Kota Bandung di wilayah Kelurahan Antapani Wetan, Bandung. Yang dijadikan sebagai variabel respons dalam data ini adalah tingkat kepuasan konsumen yang terdiri dari dari tiga kategori, yaitu Sangat Puas, puas, dan Tidak Puas. Dengan demikian, distribusi yang dipertimbangkan dalam pemodelan ini adalah distribusi multinomial.

Identifikasi Masalah

Berdasarkan uraian dari latar belakang di atas, maka masalah yang dapat diidentifikasi adalah:

1. Bagaimana mendeteksi masalah overdispersi pada data berdistribusi multinomial ?
2. Bagaimana mengatasi masalah overdispersi pada data berdistribusi multinomial dengan menggunakan model Quasi-likelihood ?

Tujuan Penelitian

Berdasarkan identifikasi masalah, maka tujuan yang ingin dicapai dalam penelitian ini adalah:

1. Untuk mendeteksi masalah overdispersi pada data yang berdistribusi multinomial.
2. Untuk mengatasi masalah overdispersi pada data yang berdistribusi multinomial dengan menggunakan model Quasi-likelihood.

B. Metodologi Penelitian

Data dapat dinyatakan mengalami masalah overdispersi ketika nilai devians atau chi kuadrat *Pearson* yang lebih dari 1 (McCullagh & Nelder, 1989). Distribusi multinomial merupakan salah

satu metode yang dapat digunakan untuk mengatasi masalah overdispersi pada data yang mengikuti distribusi normal. Model *quasi-likelihood* yang dibahas dalam skripsi ini merupakan salah satu model yang dapat digunakan untuk mengatasi masalah overdispersi pada data yang mengikuti distribusi multinomial. Data yang digunakan penulis untuk penerapan model *quasi-likelihood* adalah data tingkat kepuasan pengguna PDAM di kelurahan Antapani Wetan Bulan Oktober Tahun 2019.

C. Hasil Penelitian dan Pembahasan

Model Regresi Logistik Multinomial

Model regresi logistik multinomial pada sub bab ini di ambil dari hasil perhitungan melalui SAS yang data nya yaitu kepuasan pengguna PDAM di kelurahan Antapani Wetan, berikut ditampilkan output untuk melihat model pada tabel 4.4

Tabel 1. Penduga Parameter Regresi Logistik

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	p-value
Intercept	**1**	1	-2.5006	0.3315	56.8883	<.0001
Intercept	**2**	1	-1.1497	0.2949	15.1969	<.0001
R1W		1	0.5223	0.4235	1.5211	0.2175
R2W		1	0.6245	0.3686	2.8705	0.0902
R3W		1	0.9746	0.3833	6.4637	0.0110

Dengan demikian kita memperoleh model regresi logistik multinomial sebagai berikut:

$$g_1(x) = -2{,}5006 + 0{,}5223R1W + 0{,}6245R2W + 0{,}9746R3W$$
$$g_2(x) = -1{,}1497 + 0{,}5223R1W + 0{,}6245R2W + 0{,}9746R3W$$

Evaluasi Model Regresi Logistik Multinomial

Untuk mengevaluasi model regresi logistik multinomial ini ada tiga aspek yang diperhatikan, yaitu menguji kecocokan model, pengujian signifikansi parameter yang berada di dalam model, baik secara simultan maupun parsial.

1. Uji Kecocokan Model

   Berikut hasil uji kecocokan model pada tabel berikut:

Tabel 2. Hasil Uji Kecocokan Model

Kriteria	Nilai	DF	Value/DF	p-value
Deviance	55.8091	35	1.5945	0.0142
Pearson	42.8484	35	1.2242	0.1700

Berdasarkan hasil pengujian tersebut menujukkan bahwa diperolehnya nilai Pearson Chi-Square sebesar 42,8484 dan P-value sebesar 0,1700 yang berarti tolak H_0 pada α sebesar 0,05. Dengan demikian dapat dikatakan bahwa model yang terbentuk tidak sesuai atau dengan kata lain ada perbedaan yang nyata antara hasil observasi dengan kemungkinan prediksi model.

2. Uji Simultan

   Adapun hasil pengujian parameter yang berada di dalam model regresi logistik secara simultan disajikan pada tabel berikut:

Tabel 3. Pengujian signifikansi penduga parameter secara simultan

Test	Chi-Square	DF	p-value
Likelihood Ratio	8.0859	3	0.0443

Test	Chi-Square	DF	p-value
Score	7.6208	3	0.0545
Wald	8.0052	3	0.0459

Berdasarkan hasil di atas, diperoleh nilai statistic G atau *Likelihood Ratio* sebesar 8,0859 dengan p-value sebesar 0,0443. Nilai G tersebut lebih besar dibandingkan nilai chi-square tabel yaitu 7,8147. Maka, keputusan yang diambil adalah tolak H_0. Sehingga, dengan tingkat kepercayaan 95 persen dapat dikatakan bahwa terdapat minimal satu variabel bebas yang memengaruhi variabel tak bebas.

3. Uji Parsial

Pengujian signifikansi untuk masing-masing parameter ini menggunakan statistik uji Wald, dimana hasil pengujian tersebut disajikan pada tabel berikut ini:

Tabel 4. Penduga Parameter Regresi Logistik

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	p-value
Intercept	**1**	1	-2.5006	0.2996	69.6449	<.0001
Intercept	**2**	1	-1.1497	0.2666	18.6046	<.0001
R1W		1	0.5223	0.3827	1.8622	0.1724
R2W		1	0.6245	0.3331	3.5142	0.0608
R3W		1	0.9746	0.3464	7.9131	0.0049

Berdasarkan output di atas, dari beberapa variabel bebas R1W, R2W dan R3W memiliki nilai W^2 lebih besar dari nilai p-value yang menandakan bahwa variabel bebas tersebut secara parsial signifikan memengaruhi kepuasan terhadap pelayanan PDAM.

Variabel tersebut adalah variabel jumlah respon pada RW tertentu tentang kepuasan terhadap PDAM. Dengan begitu, menunjukkan bahwa dengan tingkat kepercayaan 95 persen dapat dikatakan bahwa tiga variabel bebas tersebut secara signifikan memengaruhi kepuasan terhadap pelayanan PDAM.

Interpretasi Model

Sebagaimana di dalam model regresi logistik biner, di dalam model regresi logistik multinomial koefisien regresinya juga diinterpretasikan sebagai odds rasio. Hasil perhitungan odds rasio untuk model regresi logistik multinomial ini disajikan pada tabel berikut ini:

Tabel 5. Nilai *odds ratio*

Effect	Point Estimate	95% Wald Confidence Limits	
R1W	1.686	0.796	3.569
R2W	1.867	0.972	3.588
R3W	2.650	1.344	5.226

Berdasarkan hasil pengolahan, didapatkan nilai *odds ratio* untuk variabel R1W adalah sebesar 1,686, R2W memiliki nilai sebesar 1,867 dan R3W memiliki nilai *odds ratio* sebesar 2,650.

Hasil Pemodelan Regresi Logistik Multinomial Berbasis Quasi-Likelihood

Model Logistik Multinomial Quasi-Likelihood

Model regresi logistik multinomial Quasi-likelihood pada sub bab ini di ambil dari hasil perhitungan melalui SAS yang data nya yaitu kepuasan pengguna PDAM di kelurahan Antapani Wetan, berikut ditampilkan output untuk melihat model pada tabel 6

Tabel 6. Penduga Parameter Regresi Logistik Berbasis Quasi-likelihood

Variable	DF	Estimate	StdErr	WaldchiSq	ProbChiSq
Intercept	1	-2,5006	0,3315	56,8883	<0,0001
Intercept	1	-1,1497	0,2949	15,1969	<0,0001
R1W	1	0,5223	0,4235	1,5211	0,2175
R2W	1	0,6245	0,3686	2,8705	0,0902
R3W	1	0,9746	0,3833	48,8808	0,0110

Dengan demikian kita memperoleh model regresi logistik multinomial quasi-likelihood sebagai berikut:

$$g_1(x) = -2{,}5006 + 0{,}5223R1W + 0{,}6245R2W + 0{,}9746R3W$$
$$g_2(x) = -1{,}1497 + 0{,}5223R1W + 0{,}6245R2W + 0{,}9746R3W$$

Evaluasi Model Regresi Logistik Multinomial Quasi-Likelihood

Berdasarkan Tabel 4.10 hasil perhitungan melalui SAS dapat dilakukan pengujian signifikansi dari model yang telah diperoleh menggunakan uji simultan dan uji parsial sebagai berikut:

Tabel 7. Pengujian Signifikansi Penduga Parameter Regresi Logistik Multinomial Quasi-Likelihood

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	8.0859	3	0.0443
Score	7.6208	3	0.0545
Wald	8.0052	3	0.0459

1. Uji Simultan
   Berdasarkan tabel 4.10 di atas, diperoleh nilai statistic G atau *Likelihood Ratio* sebesar 8,0859 dengan p-value sebesar 0,0443. Nilai G tersebut lebih besar dibandingkan nilai chi-square tabel yaitu 7,8147. Maka, keputusan yang diambil adalah tolak H_0. Sehingga, dengan tingkat kepercayaan 95 persen dapat dikatakan bahwa terdapat minimal satu variabel bebas yang memengaruhi variabel tak bebas.
2. Uji Parsial
   Pengujian signifikansi untuk masing-masing parameter ini menggunakan statistik uji Wald, Berdasarkan output di atas nilai W^2 yaitu 8,0052 lebih besar dari nilai $X^2_{(v,\propto)}$ yaitu 0,0459 yang menandakan bahwa variabel bebas tersebut secara parsial signifikan memengaruhi kepuasan terhadap pelayanan PDAM.
   Dari fungsi logit tersebut dapat diperoleh fungsi probabilitas kepuasan pengguna PDAM di kelurahan Antapani Wetan untuk yaitu sebagai berikut.

$$\pi_0(x) = \frac{1}{1 + exp\ g_1(x) + exp\ g_2(x)} = \frac{1}{1 + \exp(-0{,}3792) + \exp(0{,}9717)} = 0{,}2311$$
$$\pi_1(x) = \frac{exp\ g_1(x)}{1 + exp\ g_1(x) + exp\ g_2(x)} = \frac{\exp(-0{,}3792)}{1 + \exp(-0{,}3792) + \exp(0{,}9717)} = 0{,}0876$$
$$\pi_2(x) = \frac{exp\ g_2(x)}{1 + exp\ g_1(x) + exp\ g_2(x)} = \frac{\exp(0{,}9717)}{1 + \exp(-0{,}3792) + \exp(0{,}9717)} = 0{,}2246$$

Interpretasi Model Logistik Multinomial Quasi-Likelihood

Tabel 8. Nilai *Odds Ratio* Regresi Logistik Multinomial Quasi-likelihood

Effect	**Point Estimate**	**95% Wald Confidence Limits**	
R1W	1.686	0.796	3.569
R2W	1.867	0.972	3.588
R3W	2.650	1.344	5.226

Berdasarkan hasil pengolahan, didapatkan nilai odds ratio untuk variabel R1W adalah sebesar 1,686, R2W memiliki nilai sebesar 1,867 dan R3W memiliki nilai odds ratio sebesar 2,650.

D. Kesimpulan

Penelitian ini mendeteksi dugaan ada tidaknya masalah overdispersi dan menaksir parameter pada data berdistribusi multinomial yang terklaster. Data yang digunakan adalah "Data Tingkat Kepuasan Pengguna PDAM Kelurahan Antapani Wetan Tahun 2019" yang diolah menggunakan *software* SAS dan *Microsoft excel*. Berdasarkan penelitian dapat disimpulkan bahwa:

1. Berdasarkan pengujian diperoleh nilai skala 1,2602 yang lebih dari 1 artinya ada masalah overdispersi dalam data hal ini ditunjukkan oleh hasil pengujian hipotesis yang signifikan.
2. Data mengenai tingkat kepuasan pengguna PDAM di kelurahan Antapani Wetan tahun 2019 menujukkan bahwa diperolehnya nilai Pearson Chi-Square sebesar 42,8484 dan P-value sebesar 0,1700 yang berarti H0 ditolak pada α sebesar 0,05. Dengan demikian dapat dikatakan bahwa model yang terbentuk tidak sesuai atau dengan kata lain ada perbedaan yang nyata antara hasil observasi dengan kemungkinan prediksi model
3. Nilai parameter dari fungsi probabilitas kepuasan pengguna PDAM di kelurahan Antapani Wetan tahun 2019 untuk π_1 adalah 0,2311 artinya peluang atau proporsi dari 310 responden ada 23,11% yang merasa Sangat Puas terhadap layanan PDAM, sedangkan nilai taksiran dari parameter π_2 sebesar 0,0876 artinya peluang atau proporsi dari 310 responden ada 08,76% yang merasa puas terhadap layanan PDAM dan yang merasa Tidak Puas terhadap layanan PDAM dengan proporsi atau peluang yaitu sebesar 0,2246 atau 22,46%.

Daftar Pustaka

[1] Agresti, A. (1990). Categorical Data Analysis. New York: John Wiley and Sons.

[2] Agresti, A. (2007). An Introduction to Categorical Data Analysis (Second Edition). New York: John Wiley and Sons.

[3] Cox, David. R., Snell, E.J.; "Analysis of Binary Data", 2nd Edition, Chapman & Hallm London, 1989.

[4] D. W. Hosmer dan Lemeshow. 2000 Applied Logistic Regression. USA: John Wiley and Sons.

[5] Hajarisman, N. (2009). Analisis Data Kategorik. Bandung: Pusat Penerbitan Unisba.

[6] Hajarisman, N. (2010). Analisis Data Kategorik. Bandung: Pustaka Ceria.

[7] Hajarisman, N. (2010).Pendekatan Fungsi Quasi-Likelihood dan Implementasinya dalam Sistem SAS. Bandung: Universitas Islam Bandung

[8] McCullagh, P., and J.A. Nelder. (1989). Generalized Linear Models. (Second Edition). New York: Chapman and Hall. Wilson .

[9] Morel,Jorge G., and Neerchal, Nagaraj K. (1993). A Finite Mixture Distribution for

[10] Modelling Multinomial Extra Variation. Biometrika, 80, 363-371.
[11] Morel,Jorge G., and Neerchal, Nagaraj K. (2011). Overdispersion Models in SAS.Cary, NC, USA: SAS Institute Inc.
[12] Salim & Haidir. (2019). Penelitian Pendidikan : Metode, Pendekatan, dan Jenis. Jakarta: Kencana.
[13] Solimun. (2001). Structural Equation Modelling dan LISREL. Malang: FMIPA
[14] Universitas Brawijaya.
[15] Sugiyono. (2012). Metode Penelitian Kuantitatif Kualitatif dan R&D. Bandung: Alfabeta.
[16] Sugiyono (2015). Metode Penelitian Kombinasi (Mix Methods). Bandung: Alfabeta.
[17] Fauziah, Ghina, Sunendiari, Siti. (2021). *Estimasi Pseudo Poisson Maximum Likelihood untuk Mengatasi Masalah dalam Model Log-Linear pada Kasus Kusta di Jawa Barat Tahun 2018*. Jurnal Riset Statistika. 1(1) 57-62.