BAREKENG: Jurnal Ilmu Matematika dan Terapan
June 2021 Vol. 15 No. 2 Page 315 – 322
P-ISSN: 1978-7227 E-ISSN: 2615-3017
: https://doi.org/10.30598/barekengvol15iss2pp315-322

APLIKASI *K-FOLD CROSS VALIDATION* DALAM PENENTUAN MODEL REGRESI BINOMIAL NEGATIF TERBAIK

Application of K-fold Cross Validation in Determining the Best Negative Binomial Regression Model

Yekti Widyaningsih[1*], Graceilla Puspita Arum[2], Kevin Prawira[3]

[1,2] Prodi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Indonesia
Kampus Baru UI Depok, Jawa Barat, 16424, Indonesia

Corresponding author e-mail: [1] yekti@sci.ui.ac.id*

Abstrak

Publikasi ilmiah merupakan salah satu indikator penilaian terhadap kualitas akademisi. Tetapi tidak dapat dipungkiri pembuatan publikasi ilmiah bukanlah suatu hal yang mudah, karena membutuhkan proses pembuatan dan proses penelaahan yang rumit. Tujuan dari penelitian ini adalah untuk mengetahui faktor-faktor yang memengaruhi banyaknya publikasi ilmiah yang dihasilkan oleh mahasiswa PhD Biokimia tahun 1997. Karena variabel dependen merupakan *count data*, metode analisis yang digunakan adalah Regresi Poisson. Namun karena data mengalami overdispersi, akan digunakan Regresi Binomial Negatif. Perbandingan beberapa model Regresi Poisson dan Binomial Negatif dilakukan untuk menentukan model terbaik dengan *k-fold cross validation* sebagai validasi model. Hasil penelitian menunjukkan bahwa model terbaik yang didapatkan adalah model Regresi Binomial Negatif dengan variabel independen jenis kelamin, status pernikahan, banyaknya anak dibawah 5 tahun, prestise, dan banyaknya artikel oleh mentor dalam 3 tahun terakhir.

Kata Kunci : *Overdispersi, publikasi ilmiah, regresi binomial negatif, k-fold cross validation*

Abstract

Academic publication is an indicator for the quality of academics or the institutions. It cannot be denied that making academic publication is not an easy thing, because it requires complicated manufacturing and review process before it can be published. This study aims to determine factors that affect the number of academic publications produced by Biochemistry PhD students in 1997. Because the dependent variable is a count data, we fit Poisson Regression for the data. But since overdispersion occurred, we tried to fit Negative Binomial Regression for the data. Comparison of some Poisson and Negative Binomial conducted to determine the best model with k-fold cross validation as the validation metric. The results showed that Negative Binomial Regression with independent variable gender, marital status, number of children aged five or younger, prestige, and count of articles produced by a PhD mentor during the last three years is the best model.

Keywords: *Overdispersion, academic publication, Negative Binomial Regression, k-fold cross validation*

Article info:

Submitted: 25th January 2021 *Accepted: 03th May 2021*

How to cite this article:

Y. Widyaningsih, G. P. Arum, and K. Prawira, "APLIKASI K-FOLD CROSS VALIDATION DALAM PENENTUAN MODEL REGRESI BINOMIAL NEGATIF TERBAIK", *BAREKENG: J. Il. Mat. & Ter.*, vol. 15, no. 02, pp. 315-322, Jun. 2021.

1. PENDAHULUAN

Definisi penelitian menurut Peraturan Menteri Riset, Teknologi, dan Pendidikan Tinggi Nomor 42 Tahun 2016 tentang Pengukuran dan Penetapan Tingkat Kesiapterapan Teknologi merupakan kegiatan yang dilakukan menurut kaidah dan metode ilmiah secara sistematis untuk memperoleh informasi, data, dan keterangan yang berkaitan dengan pemahaman dan atau pengujian suatu cabang ilmu pengetahuan dan teknologi [1]. Tujuan penelitian ialah menemukan sesuatu cara atau metode tertentu dan hasilnya adalah temuan baru. Salah satu cara untuk menyampaikan temuan atau pengetahuan baru tersebut adalah melalui publikasi ilmiah.

Publikasi ilmiah merupakan salah satu kunci bagi peneliti untuk menyebarluaskan sebuah temuan baru hasil penelitian. Sebuah laporan penelitian yang hanya disimpan di perpustakaan universitas atau sebuah pusat studi hanya dapat diakses oleh kalangan yang sangat terbatas. Sementara terdapat banyak orang di seluruh dunia yang sedang mencari referensi untuk mendukung riset mereka. Publikasi ilmiah juga memberikan dampak positif bagi peneliti, institusi, dan negara [2].

Kementrian Riset dan Pendidikan Tinggi (Kemenristekdikti) mengeluarkan surat edaran (B/565/B.B1/HK.01.01/2019) terkait "Sarana Publikasi Ilmiah Mahasiswa" tertanggal 8 Juli 2019. Surat edaran ini dikeluarkan didasarkan pada Peraturan Menteri Riset, Teknologi, dan Pendidikan Tinggi Nomor 50 Tahun 2018 tentang Perubahan atas Peraturan Menteri Riset, Teknologi, dan Pendidikan Tinggi Nomor 44 Tahun 2015 tentang Standar Nasional Pendidikan Tinggi, pada bagian lampiran telah mengatur tentang karya ilmiah yang dihasilkan oleh mahasiswa berbagai Program Pendidikan dari tingkat Sarjana Terapan (D4), Sarjana (S1), Magister (S2), Magister Terapan, Doktor (S3), dan Doktor Terapan. *Doctor of Philosophy* (PhD) adalah gelar doktor yang diberikan untuk program bidang ilmu alam, teknik, dan humaniora [3].

Untuk menganalisis faktor seorang mahasiswa PhD dalam penulisan publikasi ilmiah dapat digunakan analisis regresi. Analisis regresi merupakan suatu metode yang digunakan untuk menganalisis hubungan antara variabel dependen dengan beberapa variabel independen. Analisis regresi dibedakan atas analisis regresi linear dan analisis regresi non linear. Analisis regresi linear, memiliki parameter yang linear dan menyebar normal. Apabila data dari variabel dependen yang hendak dianalisis tidak menyebar normal dan tidak linear secara parameter, maka analisis regresi yang digunakan adalah analisis regresi non linear. Memodelkan data pada analisis regresi nonlinear dapat menggunakan *Generalized Linear Model* (GLM). Salah satu analisis regresi nonlinear yang dimodelkan menggunakan GLM adalah analisis regresi Poisson. Regresi Poisson sering kali digunakan untuk menganalisis data diskrit *(count data)*. Pada penerapannya, analisis regresi Poisson harus memenuhi asumsi ekuidispersi, yaitu nilai variabel respons memiliki nilai rata-rata dan varian yang sama. Tetapi kenyataan yang ada di lapangan ialah sulit menemukan kondisi serupa, sehingga sering terjadi pelanggaran asumsi berupa overdispersi, yaitu nilai varian lebih besar dari nilai rata-rata atau underdispersi, yaitu nilai varian lebih kecil dari nilai rata-rata [4]. Terdapat banyak metode regresi yang dapat digunakan untuk mengatasi asumsi ekuidispersi yang tidak terpenuhi, salah satunya adalah regresi Binomial Negatif.

Studi ini bertujuan untuk mengetahui faktor-faktor apa saja yang mempengaruhi jumlah penulisan publikasi ilmiah yang dihasilkan oleh seorang mahasiswa PhD Biokimia tahun 1997. Data yang dipakai peneliti ialah jumlah penulisan publikasi yang dihasilkan oleh seorang mahasiswa PhD Biokimia. Data terdiri dari 915 mahasiswa sebagai sampel dengan 6 variabel yaitu, banyaknya artikel yang dihasilkan oleh mahasiswa Ph.D. dalam 3 tahun terakhir, jenis kelamin (1 untuk wanita, 0 untuk lainnya), status pernikahan (1 untuk menikah, 0 untuk lainnya), banyaknya anak dibawah 5 tahun, prestise dari program Ph.D., dan banyaknya artikel oleh mentor dalam 3 tahun terakhir.

2. METODE PENELITIAN

Seperti yang telah disebutkan pada bagian pendahuluan, ada 6 variabel pada data, yaitu *art* (jumlah artikel) sebagai variabel dependen dengan *fem* (jenis kelamin), *mar* (status pernikahan), *kid5* (banyaknya anak dibawah 5 tahun), *phd* (banyaknya artikel yang dihasilkan oleh mahasiswa Ph.D. dalam 3 tahun terakhir), dan *ment* (banyaknya artikel oleh mentor dalam 3 tahun terakhir) sebagai variabel independen. Berdasarkan data yang diperoleh pada [5,6], analisis akan dimulai menggunakan metode Regresi Poisson kemudian memeriksa terjadinya overdispersi pada Regresi Poisson. Jika terjadi overdispersi pada Regresi

Poisson, akan dilanjutkan menganalisis data menggunakan metode Regresi Binomial Negatif. Untuk mengidentifikasi kebaikan model, dilihat nilai *Pearson Chi-Square* dan *Log Likelihood Ratio*. Lalu diakhiri dengan pemilihan model terbaik dengan *k-fold cross validation* sebagai validasi model.

2.1 Regresi Binomial Negatif

Distribusi Binomial Negatif merupakan perluasan dari distribusi Poisson-Gamma yang memuat parameter dispersi k. Regresi binomial negatif merupakan suatu model regresi yang digunakan untuk menganalisis hubungan antara sebuah variabel dependen dengan satu atau lebih variabel independen yang mengalami keadaan overdispersi. Regresi Binomial Negatif biasanya digunakan untuk memodelkan data dengan variabel dependen berupa *count data* dan menggunakan *Generalized Linear Model* (GLM) dengan fungsi penghubung log [7]. Sehingga, berdasarkan [8] model regresi Binomial Negatif berganda dapat dituliskan sebagai berikut:

$$\ln(\mu_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i \qquad (1)$$
$$\mu_i = e^{(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i)}$$

Dimana μ_i adalah nilai ekspektasi dari y_i yang berdistribusi Binomial Negatif, β_0 adalah nilai konstanta, $\beta_i X_{i1}$ adalah nilai variabel independen ke-i, $\beta_k X_{ik}$ adalah nilai koefisien variabel independen, dan ϵ_i adalah nilai eror. Untuk pendugaan parameter $\beta_0, \beta_1, \dots, \beta_k$ akan digunakan *Maximum Likelihood Estimator* (MLE) dengan fungsi *Likelihood:*

$$L(\boldsymbol{\beta}, a) = \prod_{i=1}^{n} \left(\prod_{r=0}^{y_i - 1} (1 + ar) \left(\frac{1}{y_i!}\right) \left(\frac{a\mu_i}{1 + a\mu_i}\right)^{y_i} \left(\frac{1}{1 + a\mu_i}\right)^{\frac{1}{a}} \right) \qquad (2)$$

Pengujian model Regresi Binomial Negatif dapat dilakukan dengan menggunakan uji kecocokan model *(goodness of fit)* dan uji parsial [9].

2.2 *Overdispersi*

Pada regresi Poisson, terdapat asumsi yang perlu dipenuhi yaitu asumsi *ekuidispersi*. Keadaan ekuidispersi adalah keadaan dimana variabel dependen memiliki nilai *mean* dan variansi yang sama. Namun, pada praktik analisis *count data* jarang sekali ditemukan data dengan kondisi ekuidispersi, tetapi lebih sering ditemukan data yang mengalami overdispersi yaitu keadaan dimana variansi variabel dependen lebih besar daripada *mean*nya. Kemungkinan lain yang mungkin ditemui adalah keadaan underdispersi dimana variansi variabel dependen lebih kecil daripada *mean*nya.

Overdispersi atau underdispersi akan menyebabkan taksiran parameter model menjadi tidak efisien, dan jika dipaksakan tetap menggunakan regresi Poisson maka dapat memicu kesalahan interpretasi signifikansi parameter model. Untuk mendeteksi adanya overdispersi pada data, akan digunakan parameter dispersi (ϕ) dalam bentuk persamaan $Var(Y) = \phi\mu$, dan berdasarkan [10] nilai ϕ dapat diestimasi menggunakan pendekatan nilai *chi-square* dibagi dengan derajat bebasnya. Atau secara matematis dapat ditulis,

$$\phi = \frac{\chi^2}{n - p - 1} \qquad (3)$$

Jika didapatkan nilai $\phi = 1$, maka asumsi ekuidispersi terpenuhi pada data. Namun, jika didapatkan nilai $\phi > 1$, dapat disimpulkan terjadi overdispersi pada data, dan jika nilai $\phi < 1$, maka dapat dikatakan terjadi underdispersi pada data [11].

2.3 Uji Parsial

Menurut [10] pengujian secara parsial digunakan untuk mengetahui apakah variabel independen berpengaruh terhadap variabel dependen secara individual yang dihasilkan. Pengujian parameter secara parsial dilakukan menggunakan hipotesis $H_0: \beta_i = 0$ dengan $H_1: \beta_i \neq 0$ menggunakan statistik uji t yang dapat dihitung dengan,

$$t = \frac{\widehat{\boldsymbol{\beta}_\iota}}{S_e(\widehat{\boldsymbol{\beta}_\iota})} \qquad (4)$$

yang memiliki daerah penolakan $|t_{hit}| > t_{\frac{\propto}{2},\nu}$ atau tolak H_0 jika nilai signifikansi $< \alpha$.

2.4 *Goodness of Fit*

Goodness of fit ialah sebutan untuk memeriksa kesesuaian model Regresi Binomial Negatif. Uji kesesuaian *(goodness of fit)* bertujuan untuk mengambil kesimpulan tentang sebaran populasi. Uji ini didasarkan pada seberapa baik kesesuaian/kecocokan antara frekuensi pengamatan yang diperoleh data sampel dengan frekuensi harapan yang diperoleh dari distribusi yang dihipotesiskan. Terdapat beberapa ukuran *goodness of fit,* antara lain *Pearson Chi-Square*, devians, *Likelihood Ratio Test, Akaike Information Criteria (AIC),* dan *Bayesian Schwartz Criteria (BSC).* Salah satu ukuran yang paling sering digunakan untuk *goodness of fit* pada *Generalized Linear Model* (GLM) adalah *pearson chi-square* yang memiliki statistik uji [12]

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \exp(\boldsymbol{X_i}\beta))^2}{\exp(\boldsymbol{X_i}\beta)} \quad (5)$$

Hipotesis yang digunakan ialah H_0: model sesuai dengan H_1: model tidak sesuai. H_0 ditolak jika nilai statistik uji χ^2 > nilai tabel χ^2_{n-p}. P-value juga dapat digunakan untuk menentukan daerah penolakan yaitu H_0 ditolak jika nilai $p - value < \alpha$. Semakin kecil nilai *pearson chi-square* pada suatu model, maka tingkat kesalahan yang dihasilkan juga semakin kecil.

2.5 *Likelihood Ratio Test*

Keuntungan menggunakan *Maximum Likelihood Estimator* (MLE) ialah dapat digunakannya *Likelihood Ratio Test* untuk menilai kecukupan model Regresi Binomial Negatif atau *Generalized Poisson I* atas Poisson karena keduanya akan mereduksi menjadi Poisson ketika parameter dispersinya sama dengan nol. Untuk menguji kecukupan model Poisson terhadap Binomial Negatif, hipotesis yang digunakan ialah $H_0: a = 0$ (model 0 (Poisson) sama dengan model 1 (Binomial Negatif)) dengan $H_1 = a > 0$ (model 1 (Binomial Negatif) lebih baik daripada model 0 (Poisson)). *Likelihod Ratio* memiliki statistik uji T dimana,

$$T = 2(\ell_1 - \ell_0) \quad (6)$$

dimana ℓ_1 dan ℓ_0 adalah log likelihood masing-masing model pada hipotesis. Statistik uji T memiliki distribusi *Chi-Square* yang asimtotik dengan derajat bebas satu [13].

2.6 *K-Fold Cross Validation*

Sebagaimana telah dijelaskan dalam [14], *cross validation* merupakan salah satu metode dalam melakukan validasi model terbaik. Teknik ini akan menguji keefektifan dari model yang dibentuk dengan melakukan penyusunan ulang (*resampling*) pada data untuk membaginya menjadi 2 bagian yaitu data *training* dan data *testing*. Data *training* akan dipakai untuk melatih model sehingga model dapat memahami pola pada data dan untuk melakukan validasi terhadap latihan model tersebut, akan digunakan data *testing* sebagai pengujiannya.

Salah satu metode dari *cross validation* yang sering digunakan adalah *k-fold cross validation* karena metode ini secara umum akan menghasilkan model yang tidak bias. Hal ini dapat terjadi karena setiap observasi pada data memiliki kesempatan untuk menjadi data *training* ataupun data *testing*. Atau dengan kata lain, kita dapat memiliki *k subset* data untuk melatih dan mengevaluasi kinerja model.

Awalnya, metode ini akan membagi data menjadi k bagian (*folds*) yang sama besar. Nilai k dibebaskan kepada peneliti, tetapi disarankan tidak terlalu besar dan tidak terlalu kecil. Nilai k yang terlalu besar akan menghasilkan model yang tidak bias, tetapi dapat membuat variansi menjadi besar sehingga dapat memicu terjadinya *overfit*. Nilai k yang terlalu kecil akan menghasilkan model yang serupa dengan metode *cross validation* biasa yang hanya membagi data menjadi *train – test* saja (dapat memicu terjadinya bias). Nilai k yang biasa digunakan adalah $k = 5$ atau $k = 10$ [15].

Setelah data dibagi menjadi k bagian yang sama besar, k-1 bagian akan digunakan sebagai data *training* untuk melatih model dan 1 bagian yang tersisa akan digunakan sebagai data *testing* untuk validasi model. Kemudian, *mean square of error* (MSE) akan dihitung untuk melihat error pada model menggunakan rumus,

(7)

$$MSE = \sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}$$

Prosedur ini akan diulang sebanyak k kali sampai semua bagian (*fold*) menjadi data *testing*. Pada setiap iterasi, *mean square of error* akan dihitung sehingga kita akan memiliki nilai dari $MSE_1, MSE_2, \dots, MSE_k$. Untuk menentukan model mana yang terbaik, akan ditinjau berdasarkan *performance metric model* yaitu melalui rata-rata MSE yang dihasilkan pada setiap iterasi, yang dihitung dengan,

$$CV_{(k)} = \frac{1}{k}\sum_{i=i}^{k} MSE_i \tag{8}$$

Model terbaik merupakan model yang memiliki rata-rata nilai MSE yang terkecil [16].

3. HASIL DAN PEMBAHASAN

Beberapa pengujian dilakukan untuk mendapatkan model terbaik dalam menggambarkan kondisi data. Tabel 1 adalah hasil estimasi parameter untuk model Regresi Poisson.

Tabel 1. Estimasi parameter model Poisson

Variabel	Taksiran $\boldsymbol{\beta}$	Std. Error	*z-value*	*p-value*
(Intercept)	0.304617	0.102981	2.958	0.0031
fem	-0.224594	0.054613	-4.112	3.92e-05
mar	0.155243	0.061374	2.529	0.0114
kid5	-0.184883	0.040127	-4.607	4.08e-06
phd	0.0128230	0.026397	0.486	0.6271
ment	0.025543	0.002006	12.733	< 2e-16

Dengan menggunakan model Regresi Poisson, didapatkan persamaan regresi sebagai berikut:

$$\hat{\mu}_i = \exp(\mathbf{0.304617} - \mathbf{0.224594}fem + \mathbf{0.155243}mar - \mathbf{0.184883}kid5 + 0.012823phd + \mathbf{0.025543}ment) \tag{9}$$

Selanjutnya akan ditinjau analisis kasus overdispersi pada data berdasarkan perhitungan nilai ϕ menggunakan nilai statistik uji χ^2 dengan derajat bebas $n - p - 1 = 909$. Hasil perhitungan untuk nilai ϕ disajikan dalam Tabel 2 berikut,

Tabel 2. Perhitungan nilai $\boldsymbol{\phi}$

Nilai $\boldsymbol{\chi^2}$	df	Nilai $\boldsymbol{\phi}$
1662.547	909	1.828984

Tabel 2 menunjukkan hasil perhitungan ϕ sebesar 1,828984 yang mana lebih besar dari 1, sehingga disimpulkan pada data terjadi overdispersi dimana variansi variabel dependen lebih besar 1,828984 kali dari meannya. Oleh karena itu, untuk mengatasi overdispersi ini akan dicoba untuk melakukan pendekatan model Regresi Binomial Negatif. Langkah selanjutnya yaitu perlu dilakukannya penaksiran parameter untuk pemodelan Regresi Binomial Negatif dengan hasil tertera pada Tabel 3.

Tabel 3. Estimasi parameter model poisson

Variabel	Taksiran $\boldsymbol{\beta}$	Std. Error	*z-value*	*p-value*
(Intercept)	0.256144	0.137348	1.865	0.062191
fem	-0.216418	0.072636	-2.979	0.002887

Variabel	Taksiran β	Std. Error	*z-value*	*p-value*
mar	0.150489	0.082097	1.833	0.066791
kid5	-0.176415	0.052813	-3.340	0.000837
phd	0.015271	0.035873	0.426	0.670326
ment	0.029081	0.003214	9.048	< 2e-16

Hasil analisis Regresi Binomial Negatif untuk data tersebut dengan *full model* diperoleh persamaan sebagai berikut:

$$\hat{\mu}_i = \exp(0.256144 - \mathbf{0.216418}fem + 0.150489mar - \mathbf{0.176415}kid5 + 0.015271phd + \mathbf{0.029082}ment) \quad (10)$$

Selanjutnya diperlukan uji parsial untuk melihat variabel independen apa saja pada model yang berpengaruh signifikan terhadap variabel dependen. Akan digunakan metode *backward regression* dimana akan dikeluarkan variabel independen yang tidak berpengaruh signifikan terhadap variabel dependen secara satu per satu dilihat dari nilai *p-value* masing-masing variabel independen, variabel independen yang pertama dikeluarkan ialah yang memiliki *p-value* paling besar dan melebihi taraf signifikansi. Langkah ini dilakukan dari *full model* sampai didapati model berisi variabel independen yang berpengaruh signifikan terhadap variabel respon.

Pengujian pertama yaitu *full model* menunjukkan bahwa variabel *phd* memiliki *p-value* $0.670326 > 0.05$ yang menyimpulkan tidak signifikannya variabel *phd* terhadap variabel *art*. Pengujian kedua menunjukkan bahwa variabel *mar* memiliki *p-value* $0.072312 > 0.05$ yang menyimpulkan tidak signifikannya variabel *mar* terhadap variabel *art*. Pengujian ketiga menunjukkan bahwa tidak ada variabel yang memiliki *p-value* > 0.05 yang menyimpulkan sudah signifikannya variabel *fem, kid5,* dan *ment* terhadap variabel *art*.

Tabel 4. Estimasi parameter model poisson

Variabel	Taksiran β	Std. Error	*z-value*	*p-value*
(Intercept)	0.391025	0.064527	6.060	1.36e-09
fem	-0.232697	0.072184	-3.224	0.00127
kid5	-0.137754	0.048153	-2.861	0.00423
ment	0.029369	0.003116	9.425	< 2e-16

Selanjutnya akan ditaksir kembali model Regresi Binomial Negatif dengan variabel independen yang sudah signifikan terhadap variabel *art*, yaitu variabel *fem, kid5*, dan *ment*. Dari Tabel 4 diperoleh model sebagai berikut:

$$\hat{\mu}_i = \exp(\mathbf{0.391025} - \mathbf{0.232697}fem - \mathbf{0.137754}kid5 + \mathbf{0.029369}ment) \quad (11)$$

Untuk memeriksa kesesuaian model, dilakukan uji *goodness of fit*. Model yang diuji ialah *full model* Regresi Binomial Negatif dan model Regresi Binomial Negatif dengan variabel independen yang sudah signifikan terhadap variabel dependen, yaitu variabel *fem, kid5,* dan *ment*. Pengujian *goodness of fit* menggunakan statistik uji *Pearson Chi-Square* dan didapatkan hasil seperti pada Tabel 5.

Tabel 5. Pengujian *goodness of fit*

Model	Variabel independen	*p-value*
Binomial Negatif	*fem, mar, kid5, phd, ment*	0.2008164
Binomial Negatif	*fem, kid5, ment*	0.2310042

Dari Tabel 5 didapatkan nilai $p - value > \alpha = 0.05$ untuk kedua model sehingga dapat disimpulkan bahwa kedua model sudah sesuai. Selanjutnya dilakukan uji *Likelihood Ratio Test* untuk menilai kecukupan

model Regresi Binomial Negatif atas Poisson karena kedua model akan mereduksi menjadi Poisson ketika parameter dispersinya sama dengan nol. Selain kedua model tersebut, juga akan diuji dengan model Regresi Poisson untuk memastikan bahwa model Regresi Poisson kurang cocok digunakan untuk data ini.

Tabel 6. Pengujian *Likelihood Ratio Test*

Model	Variabel independen	Model 2	Variabel independen	*p-value*
Poisson	*fem, mar, kid5, phd, ment*	binomial negatif	*fem, mar, kid5, phd, ment*	2.2×10^{-16}
Binomial Negatif	*fem, mar, kid5, phd, ment*	binomial negatif	*fem, kid5, ment*	0.1816

Tabel 6 menunjukkan hasil pengujian Likelihood Ratio Test. Didapatkan hasil nilai $p - value < \alpha = 0.05$ untuk pengujian pertama sehingga dapat disimpulkan bahwa *full model* Regresi Binomial Negatif lebih sesuai dengan data daripada model Regresi Poisson. Untuk pengujian kedua didapatkan hasil nilai $p - value < \alpha = 0.05$ sehingga dapat disimpulkan bahwa *full model* Regresi Binomial Negatif lebih sesuai dengan data daripada model Regresi Binomial Negatif dengan variabel independen yang sudah signifikan terhadap variabel dependen.

Untuk memastikan model terbaik, maka akan dilakukan metode validasi model menggunakan *10-fold cross validation* pada model Regresi Poisson, model Regresi Binomial Negatif dengan seluruh variabel independen, serta model Regresi Binomial Negatif dengan variabel independen yang berpengaruh signifikan saja. Tabel 7 menunjukkan hasil rata-rata *mean square of error* pada percobaan *10-fold cross validation* pada model.

Tabel 7. Hasil pengujian *10-fold cross validation*

Model	Variabel independen	*Rata-rata MSE*
Poisson	*fem, mar, kid5, phd, ment*	3.622588
Binomial Negatif	*fem, mar, kid5, phd, ment*	**3.524407**
Binomial Negatif	*fem, kid5, ment*	3.588139

Dari Tabel 7, model yang memiliki rata-rata *mean square of error* terkecil (RMSE = 3.524407) adalah model Binomial Negatif dengan variabel independen $fem, mar, kid5, phd, ment$.

4. KESIMPULAN

Pada data terjadi overdispersi pada variabel dependen (art) dimana variansinya lebih besar daripada meannya sebesar 1.828984. Oleh karena itu, untuk mengatasi overdispersi yang ada, diajukan model Regresi Binomial Negatif. Didapatkan model regresi yang sesuai dengan data adalah,

$$\hat{\mu}_i = \exp\left(0.256144 - 0.216418fem + 0.150489mar - 0.176415kid5 + 0.015271phd + 0.029082ment\right)$$

Berdasarkan hasil uji parsial, didapatkan faktor yang berpengaruh signifikan terhadap banyaknya publikasi ilmiah oleh mahasiswa Ph.D. Biokimia tahun 1997 dalam 3 tahun terakhir ialah jenis kelamin, banyaknya anak dibawah 5 tahun, dan banyaknya artikel oleh mentor dalam 3 tahun terakhir. Dan berdasarkan uji *Likelihood Ratio Test* serta *10-fold cross validation*, model terbaik untuk menggambarkan data adalah model regresi binomial negatif dengan variabel independen jenis kelamin, status pernikahan, banyaknya anak dibawah 5 tahun, prestise, dan banyaknya artikel oleh mentor dalam 3 tahun terakhir.

DAFTAR PUSTAKA

[1] Republik Indonesia. Peraturan Menteri Riset, Teknologi, dan Pendidikan Tinggi Republik Indonesia Nomor 42 Tahun 2016 Tentang Pengukuran dan Penetapan Tingkat Kesiapterapan Teknologi.

[2] M. Setiyo, *Teknik Menyusun Manuskrip dan Publikasi Ilmiah Internasional*, Yogyakarta: Deepublish, 2017.

[3] Republik Indonesia. Peraturan Menteri Riset, Teknologi, dan Pendidikan Tinggi Republik Indonesia Nomor 50 Tahun 2018 Tentang Perubahan Atas Peraturan Menteri Riset, Teknologi, dan Pendidikan Tinggi Nomor 44 Tahun 2015 Tentang Standar Nasional Pendidikan Tinggi.
[4] R. Ruliana, P. Hendikawati, and A. Agoestanto, "Pemodelan Generalized Poisson Regression (GPR) untuk Mengatasi Pelanggaran Equidispersi pada Regresi Poisson Kasus Campak Di Kota Semarang Tahun 2013", *UNNES Journal of Mathematics*, vol. 5, no. 1, pp. 39-46, May 2016.
[5] J.S. Long, "The origins of sex differences in science", *Social Forces*, vol. 68, no. 4, pp. 1297–1315, June 1990.
[6] J.S. Long, *Regression Models for Categorical and Limited Dependent Variables*. California: Sage, 1997.
[7] N.M.R. Keswari, I.W. Sumarjaya, and N.L.P. Suciptawati, "Perbandingan Regresi Binomial Negatif Dan Regresi Generalisasi Poisson Dalam Mengatasi Overdispersi (Studi Kasus: Jumlah Tenaga Kerja Usaha Pencetak Genteng Di Br. Dukuh, Desa Pejaten)", *E-Jurnal Matematika*, vol. 3, no. 3, pp. 107-115, August 2014.
[8] R. Cahyandari, "Pengujian Overdispersi pada Model Regresi Poisson (Studi Kasus: Laka Lantas Mobil Penumpang di Provinsi Jawa Barat)", *Statistika*, vol. 14, no. 2, pp. 69-76, November 2014
[9] J.M. Hilbe, *Negative Binomial Regression, 2nd edition*. New York: Cambridge University Press, 2011.
[10] R.T. Simarmata, and D. Ispriyanti, "Penanganan Overdispersi Pada Model Regresi Poisson Menggunakan Model Regresi Binomial Negatif", *Media Statistika*, vol. 4, no. 2, pp. 95-104, December 2011.
[11] Darnah, "Mengatasi Overdispersi pada Model Regresi Poisson dengan Generalized Poisson Regression I", *Jurnal Eksponensial*, vol. 2, no. 2, pp. 5-10, November 2011.
[12] V. Eminita, A. Kurnia, and K. Sadik, "Penanganan Overdispersi Pada Pemodelan Data Cacah dengan Respon Nol Berlebih (Zero-Inflated)", *FIBONACCI : Jurnal Pendidikan Matematika dan Matematika*, vol. 5, no. 1, pp. 71-80, June 2019.
[13] P.S. Pradawati, K.G. Sukarsa, and I.G.A.M. Srinadi, "Penerapan Regresi Binomial Negatif Untuk Mengatasi Overdispersi Pada Regresi Poisson", *E-Jurnal Matermatika*, vol. 2, no. 2, pp. 6-10, May 2013.
[14] Sanjay, "Why and how to Cross Validate a Model?", *Towards Data Science*, 13 November 2018, [Online]. Tersedia: https://towardsdatascience.com/why-and-how-to-cross-validate-a-model-d6424b45261f [Diakses: 6 Januari 2021].
[15] M. Kuhn, and K. Johnson, *Applied Predictive Modeling 1st edition.* Berlin: Springer, 2013.
[16] C. Zhang, *Statistical Modeling of Count Data with Over-Dispersion or Zero-Inflation Problems*. Master [Thesis]. Montclair, NJ: Montclair State Univ., 2019. [Online]. Available: Montclair State University Digital Commons.