ANALISIS REGRESI BINOMIAL NEGATIF PADA BANYAK GOL PERTANDINGAN PIALA DUNIA FIFA

Glene Felix [1*], Daniel Laorencius [2], Christian Oloan August [3], Muhammad Adli Rahmat Solihin [4], Margareth Ravse Betari Atmojo Prabowo [5]

[1,2,3,4,5] Program Studi Ilmu Aktuaria, Departemen Matematika FMIPA Universitas Indonesia
Departemen Matematika (Gedung D) FMIPA UI, Jl. Prof. DR. Sudjono D. Pusponegoro, Pondok Cina, Kecamatan Beji, Depok, Jawa Barat 16424, Indonesia

Abstrak.

Model regresi Poisson merupakan model standar untuk data diskrit dan termasuk model regresi nonlinier dengan asumsi ekuidispersi. Jika asumsi dilanggar atau terdapat overdispersi, model regresi Poisson tidak dapat digunakan. Untuk mengatasinya, salah satu metode yang digunakan adalah dengan menggunakan model regresi Binomial Negatif. Penelitian ini menggunakan model Binomial Negatif untuk mengatasi overdispersi pada data banyak gol pertandingan piala dunia FIFA. Hasil penelitian menunjukkan regresi Binomial Negatif dengan prediktor selisih ranking FIFA kedua tim dan adanya babak penalti pada pertandingan merupakan model terbaik untuk mengatasi overdispersi dalam memprediksi banyak gol pertandingan piala dunia FIFA.

Kata Kunci: *sepak bola, regresi Poisson, regresi Binomial Negatif, Piala Dunia FIFA, Overdispersi, Excess Zero*

1. PENDAHULUAN

FIFA World Cup atau Piala Dunia FIFA merupakan sebuah ajang di mana negara yang terkualifikasi memperebutkan juara dunia sepak bola. [1] Sejarah Piala Dunia FIFA ini sudah dimulai sejak tahun 1930 dan terakhir kali diadakan pada tahun 2022 dengan Qatar sebagai tuan rumah. Tidak hanya memperebutkan untuk menjadi juara, tetapi negara-negara di dunia juga ingin menjadi negara penyelenggara Piala Dunia FIFA. [2] Alasannya karena mempunyai prospek yang bagus dari segi ekonomi di mana dengan menjadi negara penyelenggara Piala Dunia FIFA, suatu negara pasti akan menjadi pusat perhatian dunia dan banyak turis dari berbagai negara akan datang ke negara tersebut. Selain itu, prospek ini juga sangat mungkin untuk menguntungkan dalam jangka panjang di mana penonton dari setiap negara yang datang akan kembali apabila mendapatkan pengalaman yang baik saat berada di negara penyelenggara tersebut dan memberikan rekomendasi ke kerabat-kerabatnya. Hal ini akan meningkatnya kegiatan ekonomi negara tersebut dengan adanya sewa penginapan, wisata kuliner, dan biaya perjalanan dengan menggunakan infrastruktur di negara penyelenggara. [3] Ajang ini cocok bagi suatu negara untuk mempromosikan kebudayaan dan pariwisatanya sehingga pada akhirnya dapat meningkatkan kegiatan ekonomi negara.

Sepak bola merupakan olahraga yang sangat populer di dunia. [4] Keseruan permainan sepak bola terlihat saat terjadinya gol atau saat pemain mencetak skor. [5] Ditambah lagi kebanyakan negara sudah memiliki tradisi sepak bola serta memiliki pemain-pemain yang hebat. Permainannya yang menarik membuat sepak bola menjadi media hiburan bagi masyarakat dunia. Namun, keseruan dari suatu pertandingan sepak bola menjadi faktor penting yang perlu dipertimbangkan pihak penyelenggara karena jika dapat memprediksi keseruan sebuah pertandingan, pertandingan yang lebih seru bisa ditempatkan ke stadion dengan kapasitas yang lebih banyak agar memaksimalkan keuntungan. Oleh karena itu, diperlukan prediksi jumlah gol yang menjadi parameter tingkat keseruan dari suatu pertandingan agar dapat meningkatkan nilai keuntungan dari pertandingan tersebut. Data yang digunakan dalam *project* ini adalah data dari pertandingan Piala Dunia FIFA mulai dari tahun 1994 dengan melibatkan beberapa faktor yang dapat menjadi penentu banyaknya gol di setiap pertandingan. Prediksi jumlah gol yang terjadi di setiap pertandingan ini dapat dilakukan dengan menggunakan regresi Poisson. Namun, ada kemungkinan terjadi overdispersi, yaitu keadaan apabila nilai variansi dari variabel responnya lebih besar dibandingkan dengan nilai mean. Kemudian, harus diperiksa juga apakah data tersebut mengandung *excess zero.* Apabila terdapat

excess zero, maka akan digunakan *Zero-Inflated* Model dan apabila tidak terdapat *excess zero*, maka akan digunakan model Binomial Negatif.

2. METODE PENELITIAN

2.1. Deskripsi Dataset

[6] Dataset FIFA World Cup yang akan diteliti diambil dari situs Kaggle yang merupakan merupakan data seluruh pertandingan sepakbola dari tahun 1993-2022 dengan 23.921 observasi. Adapun variabel-variabel yang terdapat dalam dataset tersebut adalah seperti:

1. date = tanggal pertandingan
2. home_team = tim tuan rumah
3. away_team = tim tamu
4. home_team_continent = benua tim tuan rumah
5. away_team_continentt = benua tim tamu
6. home_team_fifa_rank = peringkat FIFA tim tuan rumah pada saat pertandingan
7. away_team_fifa_rank = peringkat FIFA tim tamu pada saat pertandingan
8. home_team_total_fifa_points = jumlah total poin FIFA tim tuan rumah pada saat pertandingan
9. away_team_total_fifa_points = jumlah total poin FIFA tim tamu pada saat pertandingan
10. home_team_score = skor tim tuan rumah penuh waktu termasuk waktu tambahan, tidak termasuk adu penalti
11. away_team_score = skor tim tamu penuh waktu termasuk waktu tambahan, tidak termasuk adu penalti
12. tournament = nama turnamen
13. city = nama kota/unit administrasi tempat pertandingan dimainkan
14. country = nama negara tempat pertandingan dimainkan
15. neutral_location = menunjukkan benar atau tidak pertandingan dimainkan di tempat netral
16. shoot_out = menunjukkan benar atau tidak pertandingan termasuk adu penalti
17. home_team_result = hasil pertandingan tim tuan rumah, termasuk adu penalti
18. home_team_goalkeeper_score = skor pertandingan FIFA dari kiper berperingkat tertinggi dari tim tuan rumah
19. away_team_goalkeeper_score = skor pertandingan FIFA dari kiper berperingkat tertinggi dari tim tamu
20. home_team_mean_defense_score = rata-rata skor permainan FIFA dari 4 pemain bertahan berperingkat tertinggi dari tim tuan rumah
21. home_team_mean_offense_score = rata-rata skor permainan FIFA dari 4 pemain lini tengah berperingkat tertinggi dari tim tuan rumah
22. home_team_mean_midfield_score = rata-rata skor permainan FIFA dari 3 pemain menyerang berperingkat tertinggi dari tim tuan rumah, termasuk pemain sayap
23. away_team_mean_defense_score = rata-rata skor permainan FIFA dari 4 pemain bertahan berperingkat tertinggi dari tim tamu
24. away_team_mean_offense_score = rata-rata skor permainan FIFA dari 4 pemain lini tengah dengan peringkat tertinggi dari tim tamu
25. away_team_mean_midfield_score = rata-rata skor permainan FIFA dari 3 pemain menyerang peringkat tertinggi dari tim tamu, termasuk pemain sayap

Akan tetapi tidak seluruh variabel di atas akan digunakan dalam penelitian ini karena terdapat terlalu banyak *missing value* dan *unique value* pada variabel tersebut.

2.2 *Data Cleaning*

Akan dilakukan modifikasi pada dataset karena terdapat beberapa variabel yang tidak relevan terhadap model regresi dengan cara mereduksi beberapa variabel untuk mempermudah proses interpretasi. Langkah pertama yang dilakukan adalah membuang observasi yang bukan merupakan pertandingan Piala Dunia FIFA karena yang difokuskan disini adalah pertandingan Piala Dunia FIFA. Langkah selanjutnya adalah menghapus beberapa kolom yang memiliki banyak *missing value* dan *unique value* seperti kolom home_team_total_fifa_points dan away_team_total_fifa_points karena sebagian besar datanya bernilai 0

disebabkan sistem poin baru diadakan mulai tahun 2011. tournament karena yang kita ambil hanya pertandingan Fifa World Cup, city dan country karena terdapat terlalu banyak kelas didalamnya (terdapat ratusan lokasi), score dan home_team_win. Sehingga kolom yang tersisa adalah kolom date, home_team, away_team, home_team_continent, away_team_continent, home_team_fifa_rank, away_team_fifa_rank, home_team_score, away_team_score, neutral_location, dan shoot_out. Dari kolom yang tersisa tersebut akan dibuat variabel baru karena kami menduga variabel baru tersebut akan berpengaruh terhadap analisis regresi kami seperti rank_diff yang merupakan selisih rank fifa dari *home team rank* dan *away team rank*, rank_avg: rata-rata *home team rank* dan *away team rank*, goals: total gol *home team score* dan *away team score* , region: apakah *continent* asal *home team* dan *away team* sama, netral: apakah bertanding di daerah netral, penalty: apakah ada babak adu penalti.

2.3 Tinjauan Pustaka

2.3.1. Model Regresi Poisson

[7] Model regresi Poisson merupakan model standar untuk data *count* dan termasuk dalam model regresi nonlinear (Cameron & Trivedi, 1998). Model ini juga dapat diaplikasikan pada model yang mengandung efek spasial.
Fungsi peluang

$$P(Y_i = y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$$

Model regresi Poisson ditulis sebagai berikut (Myers, 1990)

$$y_i = \mu_i + \epsilon_i = t_i\, exp(x_i^T \boldsymbol{\beta}) + \epsilon_i \quad i = 1, 2, \ldots, n$$

Dimana μ_i adalah rata-rata jumlah kejadian dalam periode t_i

2.3.2. Overdispersi

Regresi Poisson merupakan regresi yang menggunakan data variabel dependen yang diasumsikan berdistribusi Poisson sehingga salah satu asumsi yang harus dipenuhi adalah asumsi ekuidispersi. Sebagaimana dalam distribusi poisson, nilai mean dan variansi dari variabel respon harus bernilai sama. Namun dalam kenyataannya, asumsi tersebut sering dilanggar salah satunya adalah nilai variansi lebih besar daripada mean dari variabel respon yang disebut dengan *overdispersion*. Taksiran dispersi diukur dengan *deviance* atau Pearson's Chi-Square yang dibagi derajat bebas. Data overdispersi jika taksiran dispersi lebih besar dari 1.

2.3.3. *Excess Zero*

Masalah lain yang sering terjadi pada data berdistribusi Poisson adalah data yang banyak mengandung nilai nol atau biasa disebut dengan *excess zero*. Pada data diskrit terkadang dijumpai data dengan nilai nol pada variabel responnya. Dalam beberapa kasus nilai nol ini memiliki arti, sehingga penting untuk dimasukkan dalam analisisnya. *Excess zero* ini dapat dilihat dari proporsi nilai nol yang berlebih pada variabel responnya dibanding dengan data diskrit lainnya.

2.3.4. Model Regresi Negatif Binomial

[8] Model Regresi Negatif Binomial adalah model regresi alternatif jikalau terjadi overdispersi pada count data.
Fungsi Peluang

$$P(Y_i = y_i) = \frac{\Gamma\left(y_i + \frac{1}{k}\right)}{\Gamma\left(\frac{1}{k}\right) y_i!}\left(\frac{1}{1+k\mu_i}\right)^{\frac{1}{k}}\left(\frac{k\mu_i}{1+k\mu_i}\right)^{y_i} \text{ dengan } i = 0, 1, 2, \ldots, n$$

Pada saat $k \rightarrow 0$, maka sebaran Binomial Negatif memiliki ragam $V[Y] \rightarrow \mu$. Sebaran Binomial Negatif akan mendekati suatu sebaran Poisson yang menghasilkan rata-rata dan ragam yang sama, yaitu $E[Y] = V[Y] = \mu$. Dalam model Binomial Negatif, y_i adalah variabel yang berupa count data.

Menurut Hilbe (2011), model regresi Binomial Negatif pada umumnya menggunakan fungsi penghubung logaritma atau *log-link*, yaitu

$$ln\,\mu_i = X_i^T\beta \quad \text{dengan } i = 0, 1, 2, \ldots, n$$

Dengan $ln\,\mu_i$ dan $X_i^T\beta$ terdefinisi dalam interval $(0, \infty)$.

2.3.5. Model Regresi *Zero Inflated* Poisson (ZIP)

[9] Jansakul dan Hinde (2001) mengatakan bahwa salah satu penyebab terjadinya overdispersi adalah lebih banyak observasi bernilai nol daripada yang ditaksir untuk model Regresi Poisson. Salah satu metode analisis yang diusulkan untuk lebih banyak observasi bernilai nol daripada yang ditaksir adalah model regresi ZIP. Pada ZIP, respon $Y = Y_1, \ldots, Y_n$ independen dimana $Y_i \sim 0$ dengan probabilitas p_i dan $Y_i \sim poisson(\lambda_i)$ dengan probabilitas $1 - p_i$.

Fungsi Peluang (Jansakul & Hinde, 2004)

$$P(Y_i = y_i) = p_i + (1 - p_i)e^{-\lambda_i} \text{ saat } Y_i = 0$$

$$= (1 - p_i)\frac{e^{-\lambda_i}\lambda_i^k}{k!} \text{ saat } Y_i = k,\ k = 1, 2, \ldots, 0 \le p_i \le 1$$

Lambert dalam Jansakul & Hinde (2001) menunjukkan model gabungan untuk $\lambda = (\lambda_1, \ldots, \lambda_n)$ dan $p = (p_1, \ldots, p_n)$ sebagai berikut

$$\log(\boldsymbol{\lambda}) = \mathbf{B}\boldsymbol{\beta} \text{ dan } \text{logit}(\mathbf{p}) = \log(\mathbf{p}/(1\text{-}\mathbf{p})) = \mathbf{G}\boldsymbol{\gamma}$$

dengan **B** adalah matriks variabel prediktor, **β** dan **ɣ** adalah vektor parameter yang akan ditaksir, dan **p** adalah probabilitas observasi bernilai nol.

2.3.6. Uji AIC

[10] Uji AIC adalah metode yang dapat digunakan untuk memilih model regresi terbaik yang ditemukan oleh Akaike (Grasa, 1989). Metode ini didasarkan pada metode *maximum likelihood estimation* (MLE).

$$AIC = e^{\frac{2k}{n}\frac{\sum_{i=1}^{n}\hat{u}_i^2}{n}}$$

dengan:
k = Jumlah parameter yang diestimasi dalam model regresi
n = Jumlah observasi
u = Sisa

2.3.7 Uji BIC

[11] Saat melakukan fitting model, dimungkinkan untuk meningkatkan likelihood dengan menambahkan parameter, akan tetapi dapat mengakibatkan terjadinya overfitting. Uji BIC dapat menyelesaikan masalah ini dengan memperkenalkan istilah penalti untuk jumlah parameter dalam model.

$$BIC = ln(n)k - 2ln(\widehat{L})$$

2.3.8 Uji Pearson Chi Square

[12] Dalam pengujian ini masing-masing observasi saling bebas dengan berdistribusi chi square. Uji ini sering digunakan untuk melakukan uji *goodness of fit*.

$$\chi^2 = \sum_{i=1}^{n}\frac{(O_i - E_i)^2}{E_i}$$

3. ANALISIS DAN PEMBAHASAN

3.1. Karakteristik Data

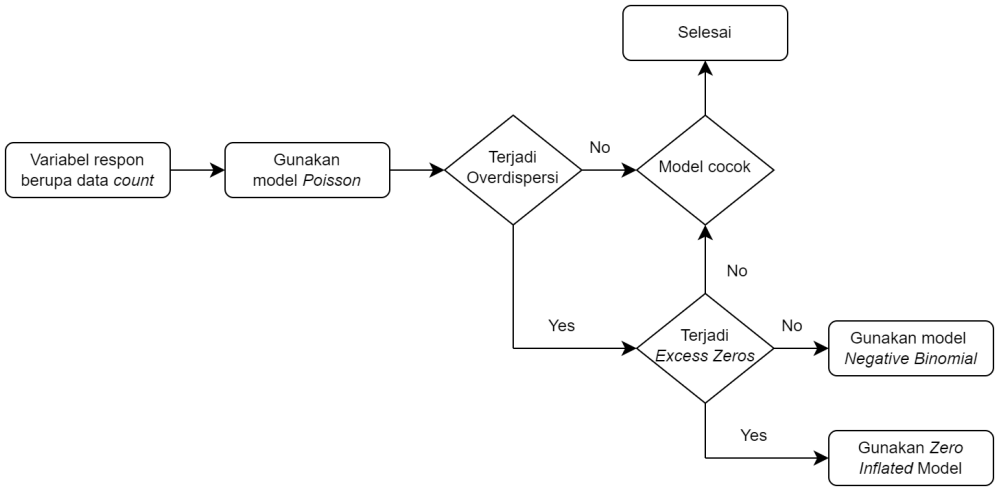
Tabel 1. Karakteristik Variabel

Variabel	Mean	Min	Max
Y	2,53700	0	8,0
X_1	0,21060	0	1,0
X_2	18,81000	1	104,0
X_3	20,65000	2	68,5
X_4	0,89810	0	1,0
X_5	0,05093	0	1,0

Sumber: Lampiran Syntax R

Tabel 1 menunjukkan karakteristik deskriptif dari variabel respon dan variabel prediktor yang digunakan pada model. Pertama, banyak gol yang tercipta dalam sebuah pertandingan memiliki rata-rata 2,54 gol dengan banyak gol berada pada rentang 0 sampai 8 gol. Kedua, rata-rata pertandingan terjadi antara dua negara yang berasal dari benua sama adalah 0,21 artinya pertandingan lebih sering terjadi dengan dua negara yang berasal dari benua berbeda. Ketiga, selisih ranking FIFA kedua tim yang bertanding memiliki rata-rata 18,81 dengan selisih ranking terkecil 1 dan terbesar 104, yang menunjukkan bahwa sebuah tim dapat menghadapi lawan yang selisih rankingnya kecil maupun besar. Keempat, rata-rata ranking FIFA kedua tim yang bertanding memiliki rata-rata 20,65 dengan rata-rata terkecilnya 2 dan rata-rata terbesarnya 68,5. Rata-rata terbesar ranking kedua tim sebesar 68,5 menunjukkan bahwa tim dengan ranking di atas 137 (68,5 dikali 2) belum pernah lolos piala dunia. Kelima, kejadian pertandingan yang terjadi di arena yang netral memiliki rata-rata 0,89 artinya pertandingan yang terjadi di daerah netral lebih sering terjadi dibandingkan di daerah non-netral. Terakhir, kejadian pertandingan dengan babak penalti memiliki rata-rata 0,05 yang menunjukkan bahwa pertandingan tanpa babak penalti lebih sering terjadi dibandingkan dengan penalti.

3.2 Menentukan Model Regresi Terbaik



Gambar 1. Diagram Alur Penentuan Model Regresi Terbaik

Bagan di atas akan menjadi acuan untuk menentukan model terbaik dari dataset ini. Karena variabel respon berupa data *count*, maka akan dimodelkan terlebih dahulu menggunakan model *Poisson*. Kemudian perlu diperiksa persebaran variansi dan mean dari variabel respon. Jika tidak mengalami overdispersi dan model yang digunakan sudah cocok dengan data, maka model regresi Poisson dapat digunakan. Namun jika mengalami overdispersi, akan kembali diperiksa apakah terjadi *excess zero* pada variabel respon. Jika terjadi overdispersi dan *excess zeros*, akan digunakan model *Zero Inflated*. Namun jika hanya terjadi overdispersi dan tidak terjadi *excess zeros*, akan digunakan model *Negative Binomial*.

3.2.1 Pengujian Model Regresi Poisson

```
Call:
glm(formula = Goals ~ Region + Rank_Diff + Rank_Avg + Netral +
    Penalty, family = poisson(link = "log"), data = data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.41955  -1.03462  -0.01613   0.42834   2.70616

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.987137   0.120145   8.216  < 2e-16 ***
Region       0.062877   0.077763   0.809  0.41876
Rank_Diff    0.005122   0.002349   2.181  0.02920 *
Rank_Avg    -0.002774   0.003368  -0.824  0.41022
Netral      -0.099983   0.099670  -1.003  0.31579
Penalty     -0.530231   0.176780  -2.999  0.00271 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 469.65  on 431  degrees of freedom
Residual deviance: 453.05  on 426  degrees of freedom
AIC: 1576.4

Number of Fisher Scoring iterations: 5
```

Gambar 2. *Summary* Regresi Poisson

Variabel respon yang digunakan pada penelitian ini merupakan variabel *count* yang menunjukkan banyak gol yang terjadi dalam sebuah pertandingan piala dunia FIFA. Karena merupakan variabel *count*, maka variabel respon diduga berdistribusi Poisson dan akan dibentuk model regresi Poisson. Dengan tingkat signifikansi $\alpha = 0,05$, hanya variabel Rank_Diff dan Penalty yang signifikan sehingga model kurang baik untuk digunakan.

Tabel 2. Nilai VIF dari Prediktor Regresi Poisson

Region	Rank_Diff	Rank_Avg	Netral	Penalty
1.109447	1.656336	1.751913	1.038206	1.029578

Sumber: Lampiran Syntax R

Selain itu, akan diperiksa juga nilai VIF dari masing-masing prediktor. Berdasarkan hasil di atas, dapat dilihat bahwa tidak terjadi multikolinearitas antar variabel karena nilai VIF dari masing-masing prediktor bernilai kurang dari 4.

3.2.2 Uji Overdispersi

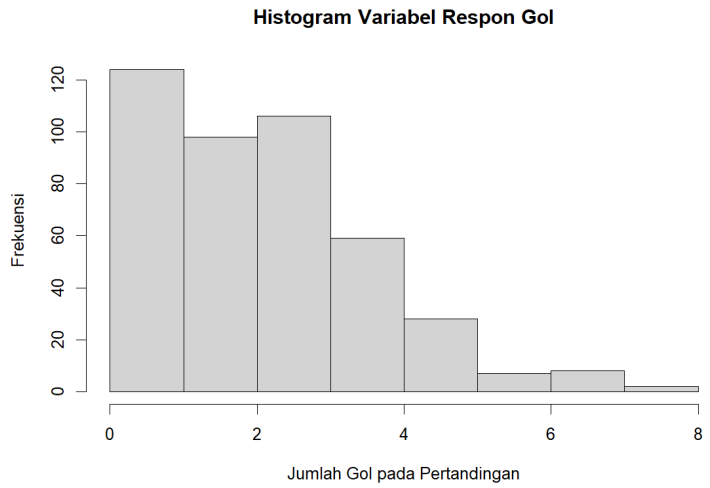
Tabel 3. Uji Pearson Chi-Square

χ-*squared*	df	χ-*squared* / df
4333.9	2155	2.01109

Sumber: Lampiran Syntax R

Untuk mengetahui apakah dataset mengalami overdispersi, maka akan dilakukan uji Chi Squared Pearson dengan melihat perbandingan nilai χ-squared dan derajat bebasnya (df). Data dikatakan mengalami overdispersi jika nilai $\frac{\chi - squared}{df} > 1$. Diperoleh nilai χ-squared = 4333,9 dan df = 2155. Karena $\frac{\chi - squared}{df} = 2,01109 > 1$, maka terjadi overdispersi. Hal ini bertentangan dengan asumsi ekuidispersi yang dimiliki model distribusi Poisson. Karena beberapa variabel prediktor tidak signifikan dan terjadi overdispersi, model regresi Poisson dianggap kurang baik dalam memodelkan data.

3.2.3 Uji *Excess Zero*



Gambar 3. Histogram Variabel Respon Gol

Selanjutnya akan diuji apakah data respon mengalami *excess zero*. Berdasarkan grafik histogram di atas, dapat dilihat bahwa frekuensi tidak terjadi gol pada pertandingan tidak berbeda drastis dengan kejadian gol lainnya, sehingga dapat disimpulkan bahwa tidak terjadi *excess zero* pada data. Karena terjadi overdispersi yang tidak disertai *excess zero*, maka diduga data berdistribusi *Negative Binomial*.

3.2.4 Pengujian Model Regresi Negative Binomial

```
Call:
glm.nb(formula = Goals ~ Region + Rank_Diff + Rank_Avg + Netral +
    Penalty, data = data, init.theta = 23488.85495, link = log)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.41947  -1.03458  -0.01613   0.42832   2.70593

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.987139   0.120151   8.216  < 2e-16 ***
Region       0.062878   0.077768   0.809  0.41879
Rank_Diff    0.005122   0.002349   2.181  0.02921 *
Rank_Avg    -0.002774   0.003369  -0.823  0.41023
Netral      -0.099985   0.099675  -1.003  0.31581
Penalty     -0.530232   0.176786  -2.999  0.00271 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(23488.85) family taken to be 1)

    Null deviance: 469.61  on 431  degrees of freedom
Residual deviance: 453.00  on 426  degrees of freedom
AIC: 1578.4
```

Gambar 4. *Summary* Regresi *Negative Binomial*

Setelah data diasumsikan berdistribusi *Negative Binomial*, akan dibentuk model regresi *Negative Binomial*. Berdasarkan gambar di atas, diperoleh fenomena yang sama seperti pada regresi Poisson, yaitu hanya prediktor Rank_Diff dan Penalty yang signifikan terhadap model dengan tingkat signifikansi $\alpha = 0,05$.

Tabel 4. Nilai VIF dari Prediktor Regresi *Negative Binomial*

Region	Rank_Diff	Rank_Avg	Netral	Penalty
1.109447	1.656336	1.751913	1.038206	1.029578

Sumber: Lampiran Syntax R

Sama halnya dengan distribusi Poisson, pada distribusi *Negative Binomial* tidak terjadi multikolinearitas antar prediktor karena nilai VIF nya yang lebih kecil dari 4. Oleh karena itu, berikutnya akan dibentuk model regresi *Negative Binomial* dengan mereduksi beberapa prediktornya.

3.2.4 Reduksi Prediktor Model Regresi *Negative Binomial*

Berikut ini akan ditinjau model regresi dengan mereduksi beberapa prediktornya selain Rank_Diff dan Penalty. Kemudian akan ditentukan model terbaiknya dengan melihat prediktor yang paling signifikan beserta nilai AIC dan BIC yang terkecil.

Model 1 adalah model tanpa mereduksi prediktornya
$\mu = exp(\beta_0 + \beta_1 Region + \beta_2 Rank\_Diff + \beta_3 Rank\_Avg + \beta_4 Netral + \beta_5 Penalty)$

Model 2 mereduksi prediktor Netral
$\mu = exp(\beta_0 + \beta_1 Region + \beta_2 Rank\_Diff + \beta_3 Rank\_Avg + \beta_5 Penalty)$

Model 3 mereduksi prediktor Rank_Avg
$\mu = exp(\beta_0 + \beta_1 Region + \beta_2 Rank\_Diff + \beta_4 Netral + \beta_5 Penalty)$

Model 4 mereduksi prediktor Region
$\mu = exp(\beta_0 + \beta_2 Rank\_Diff + \beta_3 Rank\_Avg + \beta_4 Netral + \beta_5 Penalty)$

Model 5 mereduksi prediktor Rank_Avg dan Netral
$\mu = exp(\beta_0 + \beta_1 Region + \beta_2 Rank\_Diff + \beta_5 Penalty)$

Model 6 mereduksi prediktor Region dan Netral
$\mu = exp(\beta_0 + \beta_2 Rank\_Diff + \beta_3 Rank\_Avg + \beta_5 Penalty)$

Model 7 mereduksi prediktor Region dan Rank_Avg
$\mu = exp(\beta_0 + \beta_2 Rank\_Diff + \beta_4 Netral + \beta_5 Penalty)$

Model 8 mereduksi prediktor Region, Rank_Avg, dan Netral
$\mu = exp(\beta_0 + \beta_2 Rank\_Diff + \beta_5 Penalty)$

Tabel 5. Taksiran Parameter Regresi *Negative Binomial* beserta nilai AIC dan BIC

Model	Taksiran β						AIC	BIC
	β_0	β_1	β_2	β_3	β_4	β_5		
1	0.987139***	0.062878	0.005122*	-0.002774	-0.099985	-0.530232**	1578,4	1606.882
2	0.888810***	0.065849	0.004979*	-0.002288	-	-0.509945**	1577,4	1601.796
3	0.937958***	0.076253	0.003959*	-	-0.087517	-0.528994**	1577,1	1601.499
4	1.015673***	-	0.005039*	-0.003323	-0.103095	-0.516426**	1577,1	1601.461
5	0.857723***	0.076829	0.004011*	-	-	-0.511214**	1575,9	1596.202
6	0.915564***	-	0.004890*	-0.002852	-	-0.495232**	1576.1	1596.438
7	0.961776***	-	0.003565.	-	-0.088381	-0.511016**	1576.1	1596.421
8	0.880938***	-	0.003614*	-	-	-0.493448**	1574.9	1591.14

Sumber: Lampiran Syntax R

3.3. Interpretasi Model Terbaik

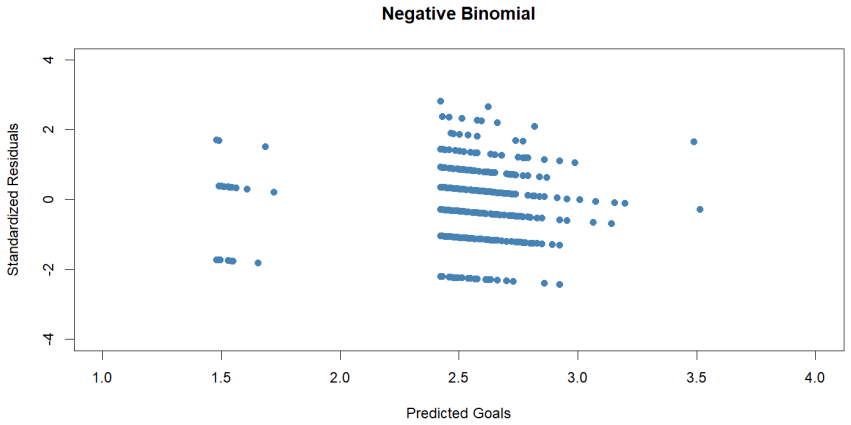
Dari hasil regresi pada Tabel 5 diperoleh Model 8 sebagai model terbaik. Hal ini dikarenakan kedua prediktornya yaitu Rank_Diff dan Penalty signifikan terhadap model. Selain itu, model ini memiliki nilai AIC dan BIC yang terkecil di antara model lainnya.

$\mu = exp(0.880938 + 0.003614\, Rank\_Diff - 0.493448\, Penalty)$

Model ini menyatakan bahwa rata-rata banyak gol akan meningkat sebesar $e^{0,003614}$ = 1,00362 kali setiap peningkatan 1 unit selisih ranking FIFA kedua tim, dengan asumsi status penalti homogen. Model ini juga menyatakan bahwa rata-rata banyak gol pertandingan dengan babak penalti akan menurun sebesar

$e^{-0,493448}$ = 0,610518 kali dibandingkan pertandingan tanpa babak penalti, dengan asumsi selisih ranking FIFA yang homogen.

3.4 Analisis Residual



Gambar 4. Plot Residual Regresi *Negative Binomial*

Meskipun tidak memiliki pola linear yang begitu jelas, dapat dilihat pada Gambar 4 bahwa plot nilai prediksi gol terhadap residualnya memiliki korelasi dengan tren yang negatif. Selain itu, jarak persebaran plotnya tidak begitu jauh sehingga model ini cocok digunakan karena nilai residualnya yang kecil. Namun terdapat kekurangan pada model ini, yaitu terjadi *gap* pada plot karena tidak mampu memprediksi nilai gol dalam tentang 1,7 – 2,4. Hal ini dikarenakan keterbatasan model dalam memprediksi rata-rata nilai gol dengan Rank_Diff yang bernilai 0, dimana tidak mungkin kedua tim bertanding dengan ranking yang sama.

3.5. Interval Kepercayaan

Tabel 6. Interval Kepercayaan Eksponensial 95%

	Estimate	2.5%	97.5%
Intercept	2.4131616	2.1952673	2.6492497
Rank_Diff	1.0036209	0.9999618	1.0071824
Penalty	0.6105177	0.4252751	0.8451342

Sumber: Lampiran Syntax R

Dengan melakukan eksponensial terhadap nilai interval kepercayaan koefisien, maka diperoleh Interval Kepercayaan 95% untuk setiap parameter pada Tabel 6.

4. KESIMPULAN

1. Rata-rata banyak gol yang tercipta dalam sebuah pertandingan adalah 2,54 gol dengan banyak gol berada pada rentang 0 sampai 8 gol. Rata-rata pertandingan terjadi antara dua negara yang berasal dari benua sama adalah 0,21 pertandingan. Rata-rata ranking FIFA kedua tim yang bertanding adalah ranking 20,65 dengan rata-rata terkecilnya 2 dan rata-rata terbesarnya 68,5. Rata-rata kejadian pertandingan yang terjadi di arena yang netral adalah 0,89 kejadian. Rata-rata kejadian pertandingan dengan babak penalti adalah 0,05 kejadian.
2. Model regresi Poisson tidak memenuhi asumsi ekuidispersi atau terjadi overdispersi dengan nilai $\frac{\chi-squared}{df} = 2,01109 > 1$ sehingga diperlukan model lain untuk memprediksi jumlah gol yang terjadi di setiap pertandingan. Model yang diusulkan adalah model regresi Binomial Negatif karena adanya overdispersi dan tidak terdapat banyak data bernilai nol atau *excess zero*.

3. Model regresi Negative Binomial terbaik adalah sebagai berikut :

$$\mu = exp(0.880938 + 0.003614\, Rank\_Diff - 0.493448\, Penalty)$$

di mana *Rank_Diff* menyatakan selisih ranking FIFA kedua tim dan *Penalty* menyatakan adanya babak penalti pada pertandingan. Model regresi Binomial Negatif menjelaskan prediksi jumlah gol yang terjadi di setiap pertandingan yang dipengaruhi oleh selisih ranking FIFA kedua tim dan adanya babak penalti pada pertandingan. Model ini menjadi model terbaik karena kedua prediktornya signifikan terhadap model dan memiliki nilai AIC dan BIC yang lebih kecil dibandingkan model regresi Binomial Negatif lainnya. Model ini menyatakan bahwa rata-rata banyak gol akan meningkat sebesar 1,00362 kali setiap peningkatan 1 unit selisih ranking FIFA kedua tim, dengan asumsi status penalty homogen. Model ini juga menyatakan bahwa rata-rata banyak gol pertandingan dengan babak penalty akan menurun sebesar 0,610518 kali dibandingkan pertandingan tanpa babak penalti, dengan asumsi selisih ranking FIFA yang homogen.

REFERENSI

[1] R. Pambudi, "Sejarah SIngkat Piala Dunia Pertama kali diadakan, Berikut Data Lengkap Juara dan Runner Up Sepanjang Masa", *iNews*, 11 Oktober 2022, [Online]. Tersedia: https://www.inews.id/sport/soccer/sejarah-singkat-piala-dunia-pertama-kali-diadakan-berikut-daftar-lengkap-juara-dan-runner-up-sepanjang-masa [Diakses: 3 Januari 2023]

[2] T. Purwanti, "Modal Piala Dunia Bikin Melongo, Rp 3.000 T. Qatar untung?", *CNBC Indonesia*, 28 November 2022, [Online]. Tersedia: https://www.cnbcindonesia.com/market/20221128115958-17-391769/modal-piala-dunia-bikin-melongo-rp-3000-t-qatar-untung#:~:text=Berdasarkan%20laporan%20dari%20Aljazeera%2C%20baru,sekitar%20Rp%20117%2C75%20triliun. [Diakses: 3 Januari 2023].

[3] S. Risanti, "Apa saja keuntungan Qatar sebagai Tuan Rumah Piala Dunia?", Fortuneidn, 25 November 2022, [Online]. Tersedia: https://www.fortuneidn.com/news/surti/keuntungan-qatar-sebagai-tuan-rumah-piala-dunia-2022 [Diakses: 3 Januari 2023]

[4] S. Wiwit, "Sepak Bola Akan Lebih Seru seandainya Banyak Gol yang Tercipta ", Kompasiana, 28 Maret 2013, [Online]. Tersedia: https://www.kompasiana.com/bunga.mawar/552e43396ea834c3338b456a/sepak-bola-akan-lebih-seru-seandainya-banyak-gol-yang-tercipta [Diakses: 3 Januari 2023]

[5] C. Matters, "Culture For All: Why Football Matters", Culture Matters, 8 April 2022, [Online]. Tersedia: https://www.culturematters.org.uk/index.php/culture/sport/item/3941-culture-for-all-why-football-matters [Diakses: 3 Januari 2023]

[6] Brenda. L, "FIFA World Cup 2022", Kaggle, 6 September 2022, [Online]. Tersedia: https://www.kaggle.com/datasets/brenda89/fifa-world-cup-2022 [Diakses: 3 Januari 2023]

[7] P. R. Sihombing, "Regresi Poisson dan Alternatifnya", Kumparan, 20 Agustus 2021, [Online]. Tersedia: https://kumparan.com/robinsihombing/regresi-poisson-dan-alternatifnya-1wMsD4Xeosa/full, [Diakses: 3 Januari 2023]

[8] D. K. Wardani, A. Wulandari, "Pemodelan Negative Binomial Regression pada Data Jumlah Kematian Bayi di Kabupaten Jombang", Tranformasi: Jurnal Pendidikan Matematika dan Matematika, vol. 4, no. 2, pp. 311-320, Desember 2020.

[9] D. Lambert, "Zero Inflated Poisson Regression, With an Application to Defects in Manufacturing", Technometrics, vol. 34, no. 1, pp. 1-14, Februari 1992.

[10] M. Fathurahman, "Pemilihan Model Regresi Terbaik dengan *Akaike's Information Criterion*", vol. 1, no. 2, September 2010.

[11] A. Datalab, "What is Bayesian Information Criterion (BIC)?", *medium*, 16 Januari 2019, [Online]. Tersedia: https://medium.com/@analyttica/what-is-bayesian-information-criterion-bic-b3396a894be6 [Diakses: 4 Januari 2023]

[12] Konsultan Data Penelitian &ArcGIS, "Mengenal Uji Pearson Chi Square Sebagai Uji Non Parametis Paling Sering Digunakan ", *patrastatistika*, 19 Agustus 2020, [Online]. Tersedia: https://patrastatistika.com/uji-pearson-chi-square/ [Diakses: 4 Januari 2023]

[13] Dobson, Annette J and Barnett, Adrian G, "An Introduction to Generalized Linear Models", ed. 3, Boca Raton, FL, USA: Chapman & Hall/CRC, 2008.

[14] W. Mendenhall, T. Sincich, "A Second Course in Statistics Regression Analysis", ed. 7, AS : Prentice-Hall, 2012.

[15] Zach, "Negative Binomial vs. Poisson: How to Choose a Regression Model", Statology, 18 Maret 2021, [Online], Tersedia: https://www.statology.org/negative-binomial-vs-poisson/, [Diakses: 3 Januari 2023]

[16] File Project Molinjut Kelompok 6 Mencakup Dataset dan Syntax R https://drive.google.com/drive/folders/1FVP_SbR-ghgPHxsajd3YqbAT7ZLmlZFL?usp=sharing