

Final Project Model Linear Lanjut **Kelompok 6**

***Analisis Regresi Negative
Binomial pada Banyak
Gol Pertandingan Piala
Dunia FIFA***

Glene Felix (2006528484) | Daniel Laorencius (2006571412) | Christian Oloan (2006571526)
Muhammad Adli (2006529184) | Margareth Ravse (2006571356)

Overview

- 01 **Abstrak**
- 02 **Pendahuluan**
- 03 **Metode Penelitian**
- 04 **Analisis dan Pembahasan**
- 05 **Kesimpulan**






Abstract

Model regresi Poisson merupakan model standar untuk data count 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.

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



Pendahuluan

- FIFA World Cup atau Piala Dunia FIFA merupakan sebuah ajang di mana negara yang terqualifikasi memperebutkan juara dunia sepak bola.
- Tidak hanya memperebutkan untuk menjadi juara, tetapi negara-negara di dunia juga ingin menjadi negara penyelenggara Piala Dunia FIFA karena mempunyai prospek yang bagus dari segi ekonomi
- 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.



Pendahuluan



- Keseruan permainan sepak bola terlihat saat terjadinya gol atau saat pemain mencetak skor.
- Namun, keseruan dari suatu pertandingan sepak bola menjadi faktor penting yang perlu dipertimbangkan pihak penyelenggara karena hal tersebut masih menjadi sebuah keraguan bagi penonton agar pertandingan yang lebih seru bisa ditempatkan ke stadion dengan kapasitas yang lebih banyak.
- Diperlukan prediksi jumlah gol yang menjadi parameter tingkat keseruan dari suatu pertandingan dan akan meningkatkan nilai keuntungan dari pertandingan tersebut.
- Data yang digunakan dalam project ini adalah data dari pertandingan Piala Dunia FIFA mulai dari tahun 1994

Deskripsi Dataset

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

- date = tanggal pertandingan
- home_team = tim tuan rumah
- away_team = tim tamu
- home_team_continent = benua tim tuan rumah
- away_team_continentt = benua tim tamu
- home_team_fifa_rank = peringkat FIFA tim tuan rumah pada saat pertandingan
- away_team_fifa_rank = peringkat FIFA tim tamu pada saat pertandingan
- home_team_total_fifa_points = jumlah total poin FIFA tim tuan rumah pada saat pertandingan
- away_team_total_fifa_points = jumlah total poin FIFA tim tamu pada saat pertandingan
- home_team_score = skor tim tuan rumah penuh waktu termasuk waktu tambahan, tidak termasuk adu penalti
- away_team_score = skor tim tamu penuh waktu termasuk waktu tambahan, tidak termasuk adu penalti
- tournament = nama turnamen
- city = nama kota/unit administrasi tempat pertandingan dimainkan

Next

Deskripsi Dataset

Lanjutan

- country = nama negara tempat pertandingan dimainkan
- neutral_location = menunjukkan benar atau tidak pertandingan dimainkan di tempat netral
- shoot_out = menunjukkan benar atau tidak pertandingan termasuk adu penalti
- home_team_result = hasil pertandingan tim tuan rumah, termasuk adu penalti
- home_team_goalkeeper_score = skor pertandingan FIFA dari kiper berperingkat tertinggi dari tim tuan rumah
- away_team_goalkeeper_score = skor pertandingan FIFA dari kiper berperingkat tertinggi dari tim tamu
- home_team_mean_defense_score = rata-rata skor permainan FIFA dari 4 pemain bertahan berperingkat tertinggi dari tim tuan rumah
- home_team_mean_offense_score = rata-rata skor permainan FIFA dari 4 pemain lini tengah berperingkat tertinggi dari tim tuan rumah
- home_team_mean_midfield_score = rata-rata skor permainan FIFA dari 3 pemain menyerang berperingkat tertinggi dari tim tuan rumah, termasuk pemain sayap
- away_team_mean_defense_score = rata-rata skor permainan FIFA dari 4 pemain bertahan berperingkat tertinggi dari tim tamu
- away_team_mean_offense_score = rata-rata skor permainan FIFA dari 4 pemain lini tengah dengan peringkat tertinggi dari tim tamu
- away_team_mean_midfield_score = rata-rata skor permainan FIFA dari 3 pemain menyerang peringkat tertinggi dari tim tamu, termasuk pemain sayap

Data Cleaning

Modifikasi pada dataset dengan cara mereduksi beberapa variabel untuk mempermudah proses interpretasi

01

Membuang observasi yang bukan merupakan pertandingan Piala Dunia FIFA

02

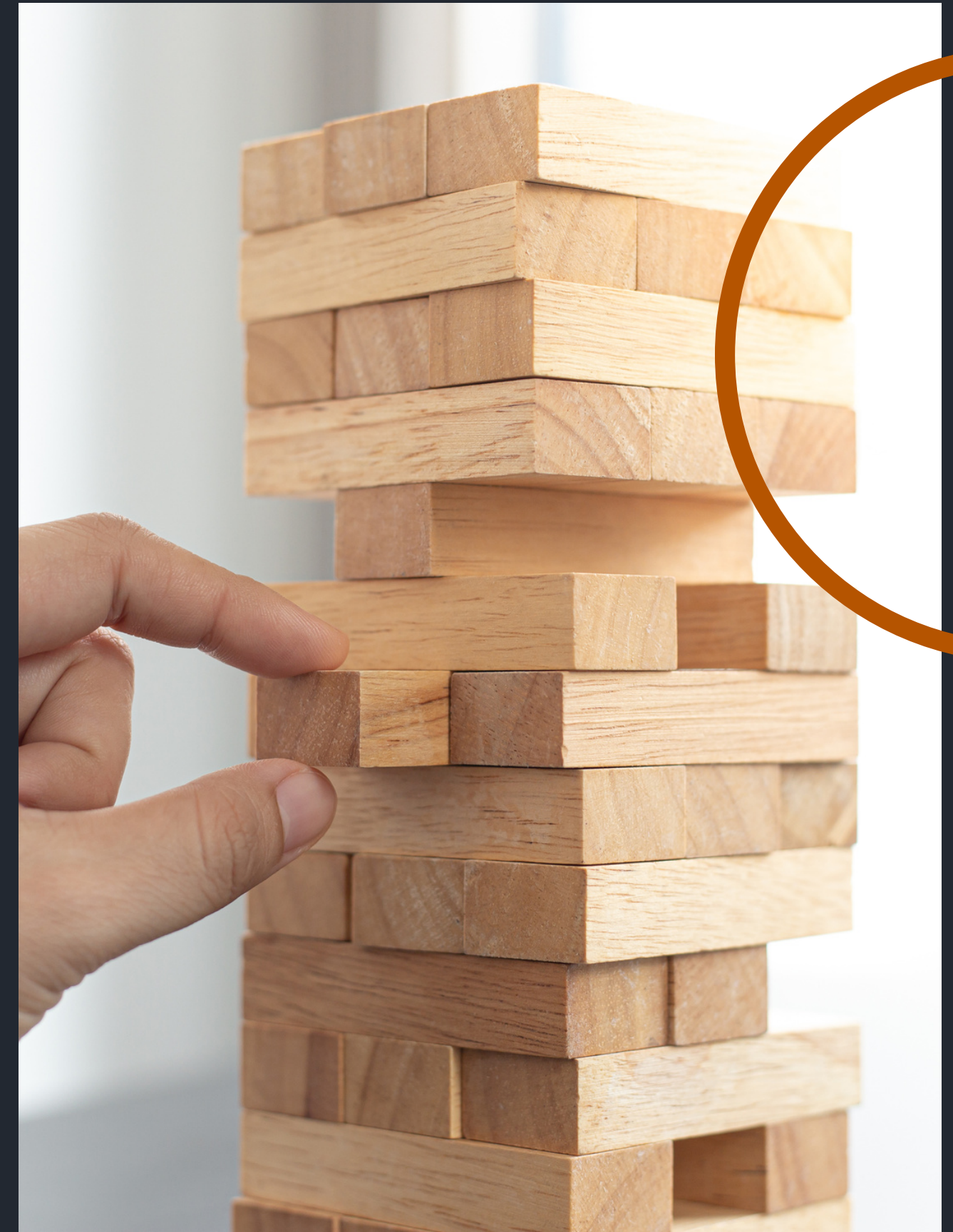
Menghapus beberapa kolom yang memiliki banyak missing value dan unique value

03

Dari kolom yang tersisa tersebut akan dibuat variabel baru karena kami menduga variabel baru tersebut akan berpengaruh terhadap analisis regresi kami

Landasan Teori

- **Model Regresi Poisson**
- **Overdispersi**
- **Excess Zero**
- **Model Regresi Negatif Binomial**
- **Model Regresi Zero Inflated Poisson**
- **Uji AIC**
- **Uji BIC**
- **Uji Pearson Chi Square**



Model Regresi Poisson

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)

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

Dimana i adalah rata-rata jumlah kejadian dalam periode t

Overdispersi



Asumsi ekuidispersi dapat dilanggar jika nilai variansinya lebih besar daripada mean dari variabel respon yang disebut dengan overdispersion. Overdispersi diukur dengan uji Pearson's Chi-Square, di mana akan terjadi overdispersi ketika Chi Square dibagi derajat bebas bernilai lebih dari 1.

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.

Model Regresi Binomial Negatif

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

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

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

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

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

Model Regresi Zero Inflated Poisson

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 $Y = Y_1, \dots, Y_n$ independen dimana $Y_i \sim 0$ dengan probabilitas p_i dan $Y_i \sim \text{poisson}(\lambda_i)$ dengan probabilitas $1 - p_i$.

Fungsi Peluang (Jansakul & Hinde, 2004)

$$\begin{aligned} P(Y_i = y_i) &= p_i + (1 - p_i)e^{-\lambda_i} \text{ saat } Y_i = 0 \\ &= (1 - p_i) \frac{e^{-\lambda_i} \lambda_i^k}{k!} \text{ saat } Y_i = k, k = 1, 2, \dots, 0 \leq p_i \leq 1 \end{aligned}$$

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

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

dengan \mathbf{B} adalah matriks variabel prediktor, $\boldsymbol{\beta}$ dan $\boldsymbol{\gamma}$ adalah vektor parameter yang akan ditaksir, dan \mathbf{p} adalah probabilitas observasi bernilai nol.

UJI AIC

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).

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

dengan:

k = Jumlah parameter yang diestimasi dalam model regresi

n = Jumlah observasi

u = Sisa

UJI BIC

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.

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

UJI Pearson Chi Square

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

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

Analisis dan Pembahasan

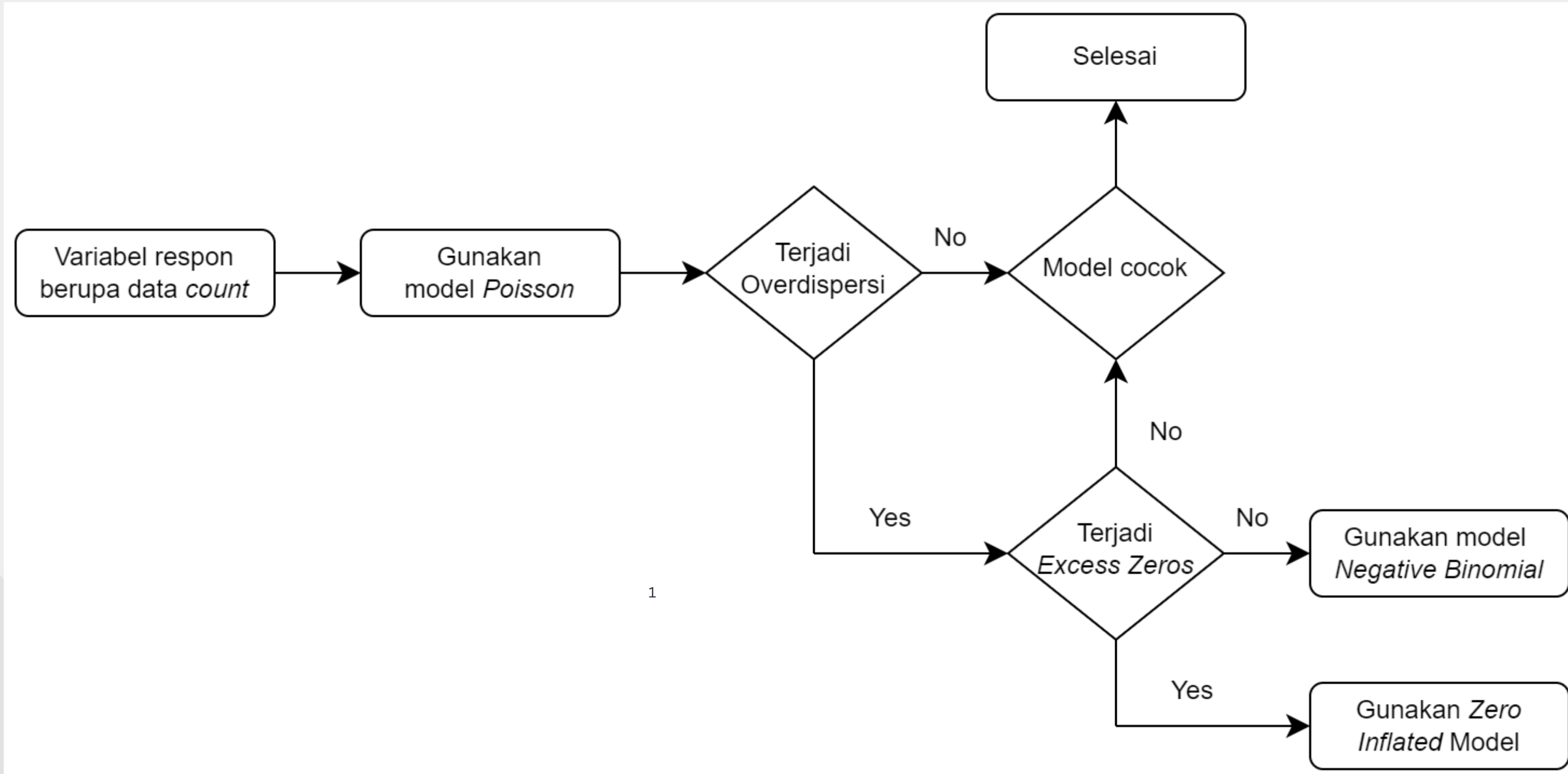


Karakteristik Data

Karakteristik Variabel			
Variabel	Mean	Min	Max
Y	2,53700	0	8,0
X1	0,21060	0	1,0
X2	18,81000	1	104,0
X3	20,65000 ¹	2	68,5
X4	0,89810	0	1,0
X5	0,05093	0	1,0

Analisis dan Pembahasan

Menentukan Model Regresi Terbaik



Analisis dan Pembahasan



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
```

```
> vif(poisson)
```

Region	Rank_Diff	Rank_Avg	Netral	Penalty
1.109447	1.656336	1.751913	1.038206	1.029578

Analisis dan Pembahasan

Uji Overdispersi



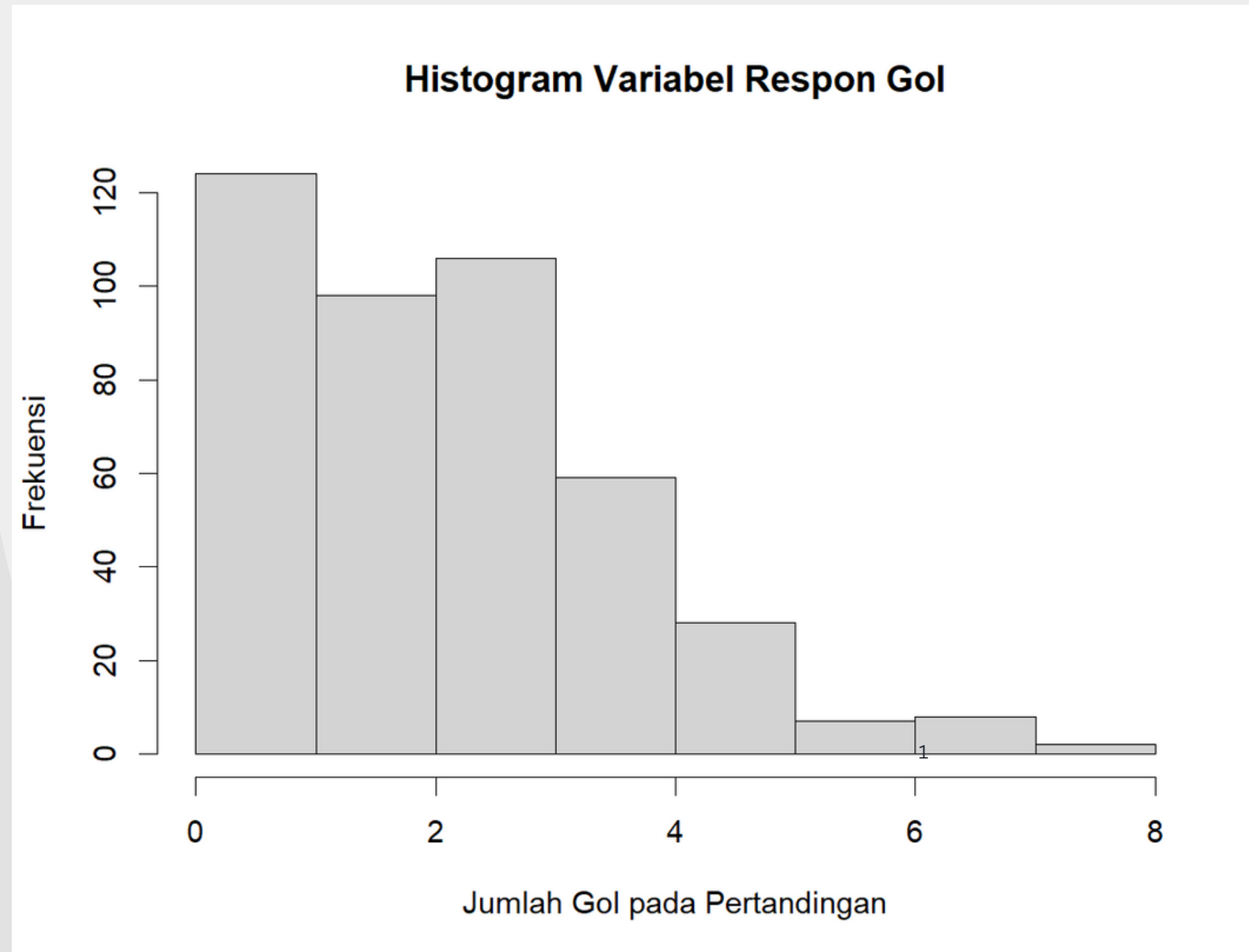
Pearson's Chi-squared test

```
data: data  
X-squared = 4333.9, df = 2155, p-value < 2.2e-16
```

**Karena $x\text{-squared}/df = 2,01109 > 1$,
maka terjadi Overdispersi**

Analisis dan Pembahasan

Uji Excess-Zero



Tidak terjadi *excess-zero* karena frekuensi banyak pertandingan dengan 0 gol tidak berbeda drastis dengan frekuensi banyak gol lainnya

Analisis dan Pembahasan



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
```

```
> vif(nb1)
```

Region	Rank_Diff	Rank_Avg	Netral	Penalty
1.109447	1.656324	1.751900	1.038206	1.029579

Analisis dan Pembahasan



Reduksi Prediktor Model Regresi Negative Binomial

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

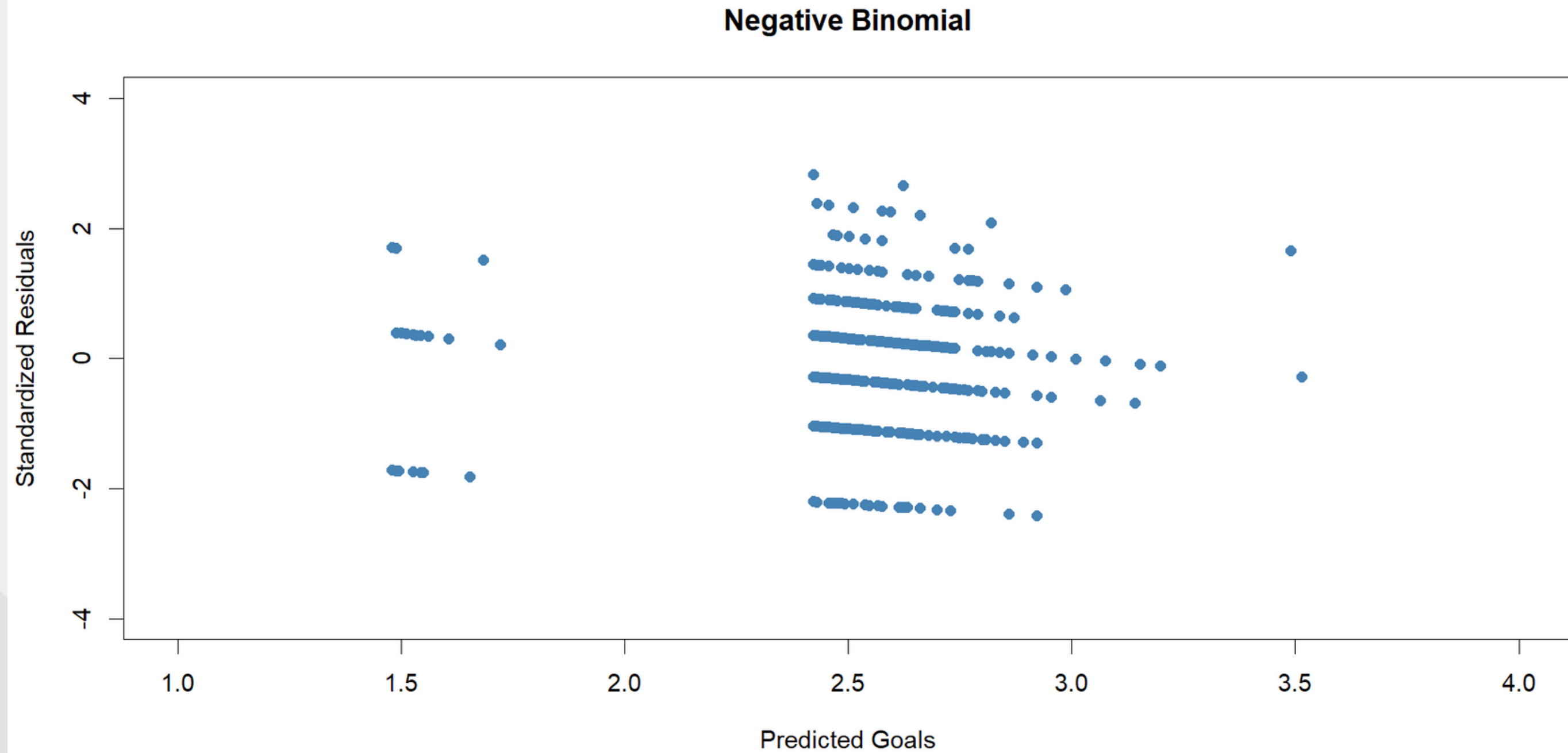
Sumber: Lampiran Syntax R

Model Terbaik

$$\mu = \exp(0.880938 + 0.003614 \text{ Rank_Diff} - 0.493448 \text{ Penalty})$$

Analisis dan Pembahasan

Analisis Residual



Analisis dan Pembahasan



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

Kesimpulan

- Rata-rata banyak gol yang tercipta dalam sebuah pertandingan adalah 2,54 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
- 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



Kesimpulan

- Model regresi Poisson tidak memenuhi asumsi ekuidispersi atau terjadi overdispersi dengan nilai :

$$\frac{\chi\text{-squared}}{df} = 2,01109 > 1$$

- Model regresi Negative Binomial terbaik :

$$\mu = \exp(0.880938 + 0.003614 \text{ Rank_Diff} - 0.493448 \text{ Penalty})$$

prediktor Rank_Diff dan Penalty signifikan terhadap model dan memiliki nilai AIC dan BIC yang lebih kecil dibandingkan model regresi Binomial Negatif lainnya

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

Presentation by **Kelompok 6**



***Thank
You!***