JURNAL MATEMATIKA, STATISTIKA & KOMPUTASI E-ISSN: 2614-8811
https://journal.unhas.ac.id/index.php/jmsk/index
Vol. 17, No. 3, 357-369, May, 2021
DOI: 10.20956/j.v17i3.12278

# Comparison of Zero Inflated Poisson (ZIP) Regression, Zero Inflated Negative Binomial Regression (ZINB) and Binomial Negative Hurdle Regression (HNB) to Model Daily Cigarette Consumption Data for Adult Population in Indonesia

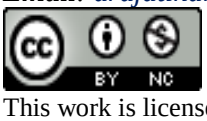
## Perbandingan Regresi *Zero Inflated Poisson* (ZIP), Regresi *Zero Inflated Negative Binomial* (ZINB) dan Regresi *Hurdle Negative Binomial* (HNB) Untuk Memodelkan Data Konsumsi Rokok Harian Penduduk Dewasa di Indonesia

**Drajat Indra Purnama ***

**Abstract**

*Smoking is a habit that is not good for health. Smoking habits are generally practiced by adults but it is possible for teenagers to do so. The Report of Southeast Asia Tobacco Control Alliance* (SEATCA) *entitled The Tobacco Control Atlas, ASEAN Region shows that Indonesia is the country with the highest number of smokers in ASEAN, namely 65.19 million people. This figure is equivalent to 34 percent of the total population of Indonesia in 2016. Based on these data, the authors are interested in modeling the daily cigarette consumption data for adults in Indonesia obtained from the 2015 Indonesia Family Life Survey. The variables used include the variable amount of cigarette consumption, education, level of welfare and income per month. The author wants to compare the best model that is most suitable for modeling the daily cigarette consumption of adults in Indonesia. The models being compared are Zero Inflated Poisson Regression (ZIP), Zero Inflated Negative Binomial Regression (ZINB) and Binomial Negative Hurdle Regression (HNB). The results of the comparison of the three models obtained that the most suitable model used for daily cigarette consumption data for adults in Indonesia is the Zero Inflated Negative Binomial (ZINB) Regression model because it has the smallest Akaike's Information Criterion (AIC) value. The results of ZINB modeling show that education, welfare and conceptual influence the daily consumption of adults in Indonesia.*

* *Badan Pusat Statistik (BPS) Kabupaten Parigi Moutong*
***Email**: drajatindrapurnama@bps.go.id*



**Keywords:** ZIP, ZINB, HNB, *Cigarette Consumption*

**Abstrak**

Merokok merupakan kebiasaan yang kurang baik bagi kesehatan. Kebiasaan merokok umumnya dilakukan oleh orang dewasa tetapi tidak menutup kemungkinan dilakukan oleh anak remaja. Laporan *Southeast Asia Tobacco Control Alliance* (SEATCA) berjudul *The Tobacco Control Atlas, ASEAN Region* menunjukkan Indonesia merupakan negara dengan jumlah perokok terbanyak di ASEAN, yakni 65,19 juta orang. Angka tersebut setara 34 persen dari total penduduk Indonesia pada 2016. Berdasarkan data tersebut, penulis tertarik untuk memodelkan data konsumsi rokok harian penduduk dewasa di Indonesia yang diperoleh dari survey Indonesia Family Life Survey 2015. Variabel yang digunakan antara lain variabel jumlah konsumsi rokok, pendidikan, tingkat kesejahteraan dan penghasilan per bulan. Penulis ingin membandingkan model terbaik yang paling sesuai digunakan untuk memodelkan konsumsi rokok harian penduduk dewasa di Indonesia. Model yang dibandingkan adalah model Regresi *Zero Inflated Poisson* (ZIP), Regresi *Zero Inflated Negative Binomial* (ZINB) dan Regresi *Hurdle Negative Binomial* (HNB). Hasil perbandingan dari ketiga model tersebut diperoleh model yang paling sesuai digunakan untuk data konsumsi rokok harian penduduk dewasa di Indonesia adalah model Regresi *Zero Inflated Negative Binomial* (ZINB) karena memiliki nilai *Akaike's Information Criterion (*AIC) paling kecil. Hasil pemodelan ZINB menunjukkan bahwa pendidikan, kesejahteraan dan penghasilan mempengaruhi konsumsi konsumsi rokok harian penduduk dewasa di Indonesia.

**Kata kunci:** ZIP, ZINB, HNB, Konsumsi Rokok

## 1. PENDAHULUAN

Merokok merupakan aktifitas membakar tembakau kemudian menghisap asapnya baik menggunakan rokok maupun pipa [2]. Pada zaman dahulu rokok berasal dari ritual pemujaan roh bangsa Indian di Amerika. Kemudian rokok dibawa oleh bangsa Eropa pada abad ke-16 sebagai pemuas kesenangan. Hingga sampai pada zaman sekarang rokok berkembang menjadi sebuah gaya hidup dan kebutuhan bagi sebagian orang.

Indonesia adalah negara dengan jumlah penduduk besar dan memiliki persentase penduduk yang merokok cukup besar. Menurut BPS [3], persentase penduduk usia 5 (lima) tahun ke atas yang merokok tembakau di Indonesia pada 2020 adalah 21,4 persen merokok setiap hari dan 1,8 persen merokok tapi tidak setiap hari dengan rata-rata jumlah batang rokok yang dihisap adalah 80,99 batang per minggu. Jika dilihat dari rata-rata pengeluaran per kapita sebulan pada 2019 untuk rokok adalah sebesar 70.537 rupiah. Pengeluaran untuk rokok ini lebih tinggi dibandingkan rata-rata pengeluaran per kapita sebulan pada 2019 untuk beras atau padi-padian sebesar 64.961 rupiah.

Terdapat beberapa faktor yang mempengaruhi perilaku merokok seseorang. Penelitian yang dilakukan Wang, *et al* [14] menggunakan data dari *China Health and Retirement Longitudinal Study* pada 2013 menunjukkan bahwa terdapat hubungan antara pendidikan dan kesejahteraan terhadap perilaku merokok. Dalam hal pendidikan, individu dengan pendidikan rendah menunjukkan perilaku merokok yang lebih tinggi daripada yang berpendidikan tinggi. Dalam hal kesejahteraan Wang, *et al* [14] membandingkan manajer dan para profesional yang hidupnya lebih sejahtera mempunyai kecenderungan untuk merokok daripada pengangguran. Kemudian

terdapat penelitian Nargis, *et al* [12] menggunakan data *Global Adult Tobacco Surveys* (2008–2011) dan *International Tobacco Control Surveys* (2009-2013) memyimpulkan bahwa penghasilan yang lebih rendah dapat membuat seseorang untuk berhenti merokok. Untuk melihat adanya pengaruh pendidikan, kesejahteraan dan penghasilan terhadap perilaku merokok diperlukan sebuah metode analisis, salah satunya adalah menggunakan analisis regresi.

Analisis regresi merupakan analisis statistika yang digunakan untuk memodelkan hubungan antara variabel respon dengan satu atau lebih variabel prediktor [8]. Hubungan antara variabel respon dan variabel prediktor dinyatakan dalam persamaan regresi. Analisis regresi umumnya digunakan untuk menganalisis data variabel respon yang berupa data kontinu. Namun dalam beberapa penelitian, variabel respon dapat berupa data diskrit. Salah satu model regresi yang dapat digunakan untuk menganalisis hubungan antara variabel respon berupa data diskrit dan variabel prediktor berupa data kontinu, diskrit atau campuran adalah model regresi Poisson.

Pada model regresi Poisson terdapat beberapa asumsi yang harus dipenuhi diantaranya adalah variabel respon berdistribusi Poisson, tidak terdapat multikolinearitas antar variabel prediktor dan *equidispersi*. *Equidispersi* adalah nilai *mean* dan varian dari variabel respon sama [11]. Pada kenyataannya terkadang tidak sepenuhnya asumsi tersebut terpenuhi, seperti nilai varian lebih besar dari nilai rata-ratanya yang disebut sebagai *overdispersi*. *Overdispersi* data *count* dapat terjadi jika mengalami *excess zero*, yaitu kondisi dimana proporsi nilai nol pada data variabel respon lebih besar dari nilai lainnya. Pada kondisi *excess zero*, regresi Poisson pada dasarnya tetap dapat diestimasi menggunakan model regresi *Zero Inflated* Poisson (ZIP). Akan tetapi pada kondisi *excess zero* dan *overdispersi*, model regresi *Zero Inflated* Poisson (ZIP) tidak sesuai untuk memodelkan data dan model yang akan terbentuk menghasilkan estimasi yang bias. Hal ini menyebabkan adanya pengembangan model-model statistik untuk mengatasi masalah *excess zero* dan *overdispersi* yaitu dengan menggunakan model regresi *Zero Inflated Negative Binomial* (ZINB) dan model regresi *Hurdle Negative Binomial* (HNB).

Permasalahan yang diangkat dalam penelitian ini adalah mencari model regresi terbaik dalam mengatasi *overdispersi dan excess zero* dengan membandingkan nilai *Akaike's Information Criterion* (AIC) [4]. Nilai AIC yang lebih kecil akan menunjukkan model yang lebih baik. Pada penelitian ini pemilihan model terbaik akan diterapkan pada data konsumsi rokok harian dari penduduk dewasa di Indonesia. Data tersebut diperoleh dari survey Indonesia Family Life Survey 2015 (www.rand.org). Akhirnya penelitian ini bertujuan untuk memperoleh model terbaik yang bisa digunakan untuk menunjukkan pola hubungan antara pendidikan, kesejahteraan dan penghasilan terhadap konsumsi rokok harian penduduk dewasa di Indonesia.

## 2. METODE PENELITIAN

### 2.1. Tahapan Penelitian

1. Melakukan analisis deskriptif variabel penelitian.
2. Melakukan identifikasi multikolinearitas.
3. Membentuk model Regresi Poisson.
4. Melakukan pengujian *overdispersi* pada *Regresi Poisson* yang terbentuk.
5. Membentuk model Regresi *Zero Inflated Poisson* (ZIP).
6. Melakukan pengujian *overdispersi* pada Regresi *Zero Inflated Poisson* (ZIP) yang terbentuk.
7. Membentuk model Regresi *Zero Inflated Negative Binomial* (ZINB).
8. Membentuk model Regresi *Hurdle Negative Binomial* (HNB).
9. Menentukan model terbaik dengan membandingkan nilai AIC.

### 2.2. Model Regresi Poisson

Model regresi Poisson merupakan penerapan dari *Generalized Linear Model* (GLM). *Generalized Linear Model* (GLM) merupakan perluasan dari model regresi yang menggambarkan hubungan antara variabel respon dengan variabel prediktor dengan variabel respon yang memiliki sebaran eksponensial. Pada regresi Poisson variabel respon menyatakan data cacah atau *count* [7]. Model regresi Poisson digunakan untuk memodelkan banyaknya kemunculan dari suatu kejadian dalam interval waktu tertentu. Menurut Myers [10], model regresi Poisson ditulis sebagai

$$y_i = \mu_i + \varepsilon_i, i = 1,2, \dots, n$$

dengan $y_i$ adalah jumlah kejadian dan $\mu_i$ adalah rata-rata jumlah kejadian yang berdistribusi Poisson. Nilai $\mu_i$ diasumsikan tetap. Nilai tengah regresi Poisson adalah $\mu_i = exp(x_i\beta)$ sehingga model regresi Poisson dituliskan sebagai

$$\ln(\mu_i) = \beta_0 + \beta_1 x_1 + \beta_1 x_1 + \cdots + \beta_p x_p$$

dengan *x* adalah variabel prediktor, $\beta$ adalah koefisien regresi, dan *p* adalah banyaknya variabel prediktor.

### 2.3. Multikolinearitas

Multikolinearitas adalah adanya korelasi linear yang sempurna atau pasti di antara beberapa atau semua variabel [6]. Jika beberapa atau semua varibel saling berkorelasi, korelasi tersebut akan menyebabkan pembobotan yang tidak berimbang sehingga akan mempengaruhi hasil analisis. Sebaiknya multikolinieritas ini tidak terjadi atau tidak terdapat multikolinieritas diantara variabel. Menurut Gujarati [6], salah satu cara indentifikasi adanya multikolinieritas adalah menghitung nilai *Varians Inflation Factor* (*VIF* ) yang dirumuskan sebagai

$$VIF = \frac{1}{1 - R^2}$$

dengan $R^2$ adalah nilai koefisien determinasi variabel respon dengan variabel prediktor. Multikolinieritas terindikasi apabila nilai VIF > 10.

### 2.4. *Overdispersi*

Suatu ciri dari distribusi Poisson adalah adanya *equidispersi*, yakni keadaan dimana nilai *mean* dan varian dari variabel respon bernilai sama. Namun kadang-kadang ditemukan keadaan yang disebut *overdispersi* yaitu varian dari variabel respon lebih besar dari rata-ratanya. Menurut Hilbe [7], salah satu cara untuk mendeteksi *overdispersi* adalah dengan *Pearson Chi-Square* dibagi dengan derajat bebas, yang dituliskan

$$\phi = \frac{\chi^2}{df} \; ; \; \chi^2 = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{var(y_i)}$$

dengan $df = n - k$ dengan $k$ merupakan banyaknya parameter termasuk konstanta, $n$ merupakan banyaknya pengamatan dan $\chi^2$ adalah *Pearson Chi-Square*. Jika $\phi > 1$ maka terjadi *overdispersi*.

### 2.5. Model Regresi *Zero Inflated* Poisson (ZIP)

Menurut Jansakul dan Hinde [9] menyatakan bahwa $y_i$ adalah variabel acak independen yang berdistribusi ZIP, maka nilai nol diasumsikan muncul dari dua yang sesuai dengan kejadian yang terpisah. Langkah pertama adalah kejadian pada probabilitas yang hanya menghasilkan observasi dengan nilai nol ($\pi_i$). Langkah kedua adalah kejadian pada probabilitas hasil

penghitungan data sebaran Poisson dengan parameter λ $(1-\pi_i)$. Dua langkah tersebut sering dituliskan

$$Y_i \sim \begin{cases} 0 & , \text{dengan probablitas } \pi_i \\ Poisson(\pi_i) & , \text{dengan probablitas } (1-\pi_i) \end{cases}$$

Sehingga Model ZIP dapat dirumuskan sebagai berikut [1]

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)e^{-\lambda_i}, \text{untuk } y_i = 0 \\ (1-\pi_i)\dfrac{e^{-\lambda_i}\lambda_i{}^{y_i}}{y_i!}, \text{untuk } y_i > 0 \end{cases}$$

dengan $\lambda_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$ dan $\pi_i = \frac{e^{\mathbf{x}_i^T\gamma}}{1+e^{\mathbf{x}_i^T\gamma}}$. Sehingga pada model regresi ZIP diperoleh model hubungan untuk $\boldsymbol{\lambda}$ dan $\boldsymbol{\pi}$ adalah

$$\log(\boldsymbol{\lambda}) = \boldsymbol{X\beta}$$
$$\text{logit}(\boldsymbol{\pi}) = \log\left(\frac{\boldsymbol{\pi}}{1-\boldsymbol{\pi}}\right) = \boldsymbol{X\gamma}$$

dengan $\boldsymbol{X}$ adalah matriks variabel prediktor, $\boldsymbol{\beta}$ adalah vektor dari parameter regresi yang diperkirakan merupakan variabel yang mempengaruhi keadaan Poisson dan $\boldsymbol{\gamma}$ adalah vektor parameter regresi yang diperkirakan merupakan variabel yang mempengaruhi keadaan nol.

### 2.6. Model Regresi Zero *Inflated Negative Binomial* (ZINB)

Model ZINB mengasumsikan terdapat dua proses pendugaan data yang berbeda yang ditentukan menggunakan uji coba *Bernoulli* [5]. Jika $y_i$ adalah variabel acak independen yang diskrit, maka nilai nol diasumsikan muncul dari dua proses terpisah, Proses pertama adalah hitungan nol dengan probabilitas $\pi_i$. Proses kedua adalah hitungan nol yang diatur oleh *Negative Binomial* dengan mean λ dan memiliki probabilitas $(1-\pi_i)$. Kemungkinan keseluruhan hitungaan nol adalah probabilitas gabungan nol dari dua proses yersebut. Jadi, model ZINB untuk respon tersebut $Y_i$ dapat ditulis sebagai

$$P(Y_i = y_i) = \begin{cases} \pi_i + (1-\pi_i)(1+k\lambda)^{-\frac{1}{k}}, \text{untuk } y_i = 0 \\ (1-\pi_i)\dfrac{\Gamma\left(y_i+\frac{1}{k}\right)(k\lambda)^{y_i}}{\Gamma(y_i+1)\Gamma\left(\frac{1}{k}\right)(1+k\lambda)^{y+\frac{1}{k}}}, \text{untuk } y_i = 1,2,\ldots \end{cases}$$

Terdapat dua komponen model regresi ZINB yaitu model data diskrit untuk $\mu_i$ dan model *zero-inflation* untuk $p_i$ yang dituliskan

$$\ln(\lambda) = X_i^T\beta$$
$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = Z_i^T\boldsymbol{\gamma}$$

### 2.7. Model Regresi *Hurdle*

Selain menggunakan model regresi ZIP, permasalahan *excess zeros* pada regresi Poisson bisa diatasi dengan model regresi *Hurdle*. Model ini adalah bagian dari model *binary* dengan variabel respon bernilai nol atau bilangan bulat positif. Berbeda dengan model regresi ZIP, model regresi

*Hurdle* menangani *outcome* dari respon secara terpisah [15]. Ketika *outcome* dari variabel respon positif, maka digunakan distribusi yang terpenggal (*truncated).*

Misalkan $P(y_i > 0) = \pi_i$ dan $P(y_i = 0) = (1 - \pi_i)$, sehingga$P(y_i|y_i > 0)$ akan mengikuti distribusi terpenggal dengan fungsi peluang $f(y_i; \mu_i)$, misalnya distribusi Poisson terpenggal. Adapun bentuk distribusi secara lengkap dapat dituliskan sebagai

$$P(Y_i = y_i) = \begin{cases} (1 - \pi_i), \text{untuk } y_i = 0 \\ \pi_i \left( \dfrac{f(j; \mu_i)}{1 - f(0; \mu_i)} \right), \text{untuk } y_i = 1,2, \dots \end{cases}$$

Model regresi HNB dibagi menjadi dua komponen yaitu model regresi logistik untuk $\pi_i$ dan model loglinier untuk $\mu_i$ dari distribusi yang tidak terpenggal yaitu

$$logit(\pi_i) = \boldsymbol{x}_{1i}\beta_1$$
$$log(\mu_i) = \boldsymbol{x}_{2i}\beta_2$$

### 2.8. Pemilihan Model Terbaik

Terdapat beberapa metode yang digunakan dalam me-nentukan model terbaik, salah satunya adalah *Akaike's Information Criterion (*AIC). Menurut Osuji, *et al* [13], AIC didefinisikan sebagai

$$AIC = -2 \ln L(\hat{\beta}) + 2p$$

Dengan $L(\hat{\beta})$ adalah nilai *likelihood* dan *p* adalah jumlah parameter. Model terbaik adalah model yang mempunyai nilai AIC terkecil

## 3. HASIL DAN PEMBAHASAN

### 3.1. Analisis Deskriptif Variabel Penelitian

Data pada penelitian ini merupakan data sekunder mengenai konsumsi rokok harian dari penduduk dewasa di Indonesia. Data tersebut diperoleh dari survey Indonesia Family Life Survey 2015 (www.rand.org). Jumlah penduduk dewasa yang menjadi objek pengamatan adalah 6.213 Orang. Terdapat 2.813 Orang (45,27 persen) yang mengkonsumsi rokok dan 3.400 Orang (54,73 persen) yang tidak megkonsumsi rokok. Variabel respon penelitian ini adalah jumlah konsumsi rokok (*Y*) sedangkan variabel prediktor adalah pendidikan ($X_1$), tingkat kesejahteraan ($X_2$), dan penghasilan per bulan ($X_3$). Data variabel penelitian disajikan pada Tabel 1.

**Tabel 1.** Variabel Penelitian

| Variabel | Deskripsi | Keterangan |
|---|---|---|
| *Y* | Jumlah Konsumsi Rokok | Batang/Hari |
| $X_1$ | Pendidikan | 1 = SD |
| | | 2 = SMP |
| | | 3 = SMA |
| | | 4 > SMA |
| $X_2$ | Kesejahteraan | 1 = Sangat miskin |
| | | 2 = Miskin |
| | | 3 = Cukup Miskin |
| | | 4 = Cukup Kaya |
| | | 5 = Kaya |
| | | 6 = Sangat Kaya |
| $X_3$ | Penghasilan | Rupiah/Bulan |

Selanjutnya dilakukan analisis deskriptif terhadap variabel prediktor. Hasil analisis deskripsi disajikan pada Tabel 2. Pada Tabel 2 terlihat bahwa nilai standar deviasi variabel penghasilan ($X_3$) memiliki nilai yang jauh berbeda dengan variabel pendidikan ($X_1$) dan variabel kesejahteraan ($X_2$), sehingga dilakukan transformasi log pada variabel penghasilan ($X_3$).

**Tabel 2.** Statistik Deskriptif Variabel Prediktor

| Keterangan | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| Minimum | 1 | 1 | 2.500 |
| Rata-rata | 2,838 | 3,068 | 2.034.141 |
| Maksimum | 4 | 6 | 999.999.997 |
| Standar Deviasi | 0,997 | 0.889 | 12.994.470 |

### 3.2. Identifikasi Multikolinearitas

Pada variabel prediktor yang digunakan dilakukan penghitungan nilai *Variance Inflation Factor* (VIF). Nilai VIF ini digunakan untuk mengecek ada tidaknya multikolinieritas. Nilai VIF $\geq$ 10 menyatakan bahwa terjadi multikolineritas. Pada Tabel 3 menunjukkan nilai VIF dari masing-masing variabel prediktor kurang dari 10, yang berarti tidak terjadi multikolinieritas.

**Tabel 3.** Nilai VIF Variabel Prediktor

| Variabel Prediktor | VIF |
|---|---|
| $X_1$ | 1,117 |
| $X_2$ | 1,102 |
| Log $X_3$ | 1,055 |

### 3.3. Regresi Poisson

Berikut merupakan penaksiran parameter model regresi Poisson yang disajikan pada Tabel 4.

**Tabel 4**. Estimasi Parameter Model Regresi Poisson

| Variabel | Estimasi | Satndar Error | Nilai Z | *p-value* |
|---|---|---|---|---|
| Intercept | -1,148128 | 0,075200 | -15,27 | < 2e-16 |
| $X_1$ | -0,234172 | 0,005559 | -42,13 | < 2e-16 |
| $X_2$ | -0,191196 | 0,006273 | -30,48 | < 2e-16 |
| Log $X_3$ | 0,290007 | 0,005499 | 52,74 | < 2e-16 |
| AIC = 74.824 | | | | |
| Devian = 63.178 | | | | |
| LRT = 5.209,3 | | | | |

Pengujian parameter regresi Poisson secara simultan didasarkan hasil pada Tabel 4. Dari Tabel 4 dapat disimpulkan bahwa $H_0$ ditolak karena nilai *Likelihood Ratio Test* (LRT) sebesar 5.209,3 lebih besar dari $\chi^2_{0,05;4} = 9{,}488$. Artinya bahwa terdapat salah satu variabel prediktor yang berpengaruh signifikan terhadap variabel respon. Berdasarkan pengujian secara parsial dengan $\alpha = 0{,}05$ didapatkan bahwa semua parameter memiliki nilai *p-value* < $\alpha = 0{,}05$. Hal ini menunjukkan bahwa semua variabel prediktor dalam model secara parsial berpengaruh secara signifikan terhadap jumlah konsumsi rokok ($Y$).

Untuk menguji asumsi *overdispersi* pada regresi Poisson dilakukan uji dengan cara membagi nilai *Pearson Chi-Square* (Devian) dengan derajat bebasnya. Jika nilai tersebut lebih dari 1, maka dikatakan terjadi overdispersi. Perhitungan *Pearson Chi-Square* (Devian) dibagi dengan derajat bebasnya adalah

$$\hat{\phi} = \frac{63.178}{6.209} = 10{,}175$$

Karena nilai rasio dispersi $\hat{\phi} = 10{,}175 > 1$, maka dapat dikatakan bahwa pada regresi Poisson terjadi *overdispersi*.

### 3.4. Model Regresi *Zero Inflated* Poisson (ZIP)

Model regresi *Zero Inflated* Poisson (ZIP) terbagi menjadi dua yaitu model *count* atau model log yang digunakan untuk menentukan peluang dari variabel respon suatu pengamatan bernilai selain nol dan model *zero inflation* atau model logit yang digunakan untuk menentukan peluang dari variabel respon suatu pengamatan bernilai nol. Pada Tabel 5 disajikan estimasi dan pengujian parameter regresi ZIP.

**Tabel 5.** Estimasi dan Pengujian Parameter Regresi ZIP

| Variabel | Estimasi | Satndar Error | Nilai Z | *p-value* |
|---|---|---|---|---|
| Model *Count* atau Model Log | | | | |
| Intercept | 0,8843966 | 0,0811169 | 10,903 | < 2e-16 |
| $X_1$ | -0,0086541 | 0,0058340 | -1,483 | 0,138 |
| $X_2$ | -0,0003443 | 0,0061519 | -0,056 | 0,955 |
| Log $X_3$ | 0,1167051 | 0,0059484 | 19,619 | < 2e-16 |
| Model *Zero Inflation* atau Model Logit | | | | |
| Intercept | 2,67841 | 0,35435 | 7,559 | 4,07e-14 |
| $X_1$ | 0,46556 | 0,02915 | 15,969 | < 2e-16 |
| $X_2$ | 0,39948 | 0,03260 | 12,254 | < 2e-16 |
| Log $X_3$ | -0,36058 | 0,02658 | -13,568 | < 2e-16 |
| AIC = 32.919,8 | | | | |
| Log *Likelihood* = -1,645e+04 | | | | |
| Devian = 13.335 | | | | |
| df = 8 | | | | |
| LRT = 1.040,7 | | | | |

Pengujian parameter regresi ZIP secara simultan didasarkan hasil pada Tabel 5. Dari Tabel 5 dapat disimpulkan bahwa $H_0$ ditolak karena nilai *Likelihood Ratio Test* (LRT) sebesar 1.040,7 lebih besar dari $\chi^2_{0,05;8} = 15{,}507$. Artinya bahwa terdapat salah satu variabel prediktor yang berpengaruh signifikan terhadap variabel respon. Berdasarkan pengujian secara parsial masing-masing variabel prediktor diperoleh

1. Pada model *count* dengan $\alpha = 0{,}05$ diperoleh nilai *p-value* < $\alpha = 0{,}05$ hanya terdapat pada variabel Log $X_3$ berarti bahwa variabel variabel Log $X_3$ berpengaruh secara signifikan terhadap jumlah konsumsi rokok ($Y$). Sedangkan untuk variabel $X_1$ dan $X_2$ tidak berpengaruh secara signifikan terhadap jumlah konsumsi rokok ($Y$).
2. Pada model *zero inflation* dengan $\alpha = 0{,}05$ diperoleh semua nilai *p-value* < $\alpha = 0{,}05$ yang berarti bahwa semua variabel prediktor berpengaruh secara signifikan terhadap jumlah konsumsi rokok ($Y$).

Untuk menguji asumsi *overdispersi* pada regresi ZIP dilakukan uji dengan cara membagi nilai *Pearson Chi-Square* (Devian) dengan derajat bebasnya. Jika nilai tersebut lebih dari 1, maka dikatakan terjadi *overdispersi*. Perhitungan *Pearson Chi-Square* (Devian) dibagi dengan derajat bebasnya adalah

$$\hat{\phi} = \frac{13.335}{2.809} = 4{,}747$$

Karena nilai rasio dispersi $\hat{\phi} = 4{,}747 > 1$, maka dapat dikatakan bahwa pada regresi ZIP terjadi *overdispersi*. Untuk mengatasi *overdispersi* pada regresi ZIP digunakan regresi *Zero Inflated Negative Binomial* (ZINB) dan regresi *Hurdle Negative Binomial* (HNB).

### 3.5. Model Regresi *Zero Inflated Negative Binomial* (ZINB)

Model regresi *Zero Inflated Negative Binomial* (ZINB) digunakan untuk menangangi overdispersi pada model ZIP. Model regresi ZINB dibagi menjadi dua komponen, yaitu model *count* atau model log dan model *zero inflation* atau model logit. Pada Tabel 6 disajikan estimasi dan pengujian parameter regresi ZINB.

**Tabel 6.** Estimasi dan Pengujian Parameter Regresi ZINB

| Variabel | Estimasi | Satndar Error | Nilai Z | *p-value* |
|---|---|---|---|---|
| Model *Count* atau Model Log | | | | |
| Intercept | 0,8614080 | 0,1774622 | 4,854 | 1,21e-06 |
| $X_1$ | -0,0059629 | 0,0130242 | -0,458 | 0,647 |
| $X_2$ | 0,0008181 | 0,0136422 | 0,060 | 0,952 |
| Log $X_3$ | 0,1170437 | 0,0129442 | 9,042 | < 2e-16 |
| Log (theta) | 1,1124197 | 0,0363070 | 30,639 | < 2e-16 |
| Model *Zero Inflation* atau Model Logit | | | | |
| Intercept | 2,60871 | 0,35758 | 7,296 | 2,97e-13 |
| $X_1$ | 0,47032 | 0,02950 | 15,944 | < 2e-16 |
| $X_2$ | 0,40274 | 0,03292 | 12,235 | < 2e-16 |
| Log $X_3$ | -0,35844 | 0,02680 | -13,373 | < 2e-16 |
| AIC = 26.656,05 | | Log *Likelihood* = -1,332e+04 | | |
| $\theta = 3{,}0417$ | | LRT = 710,83 | | |

Pengujian parameter regresi ZINB didasarkan hasil pada Tabel 6. Menggunakan $H_0$ adalah variabel prediktor tidak berpengaruh signifikan terhadap variabel respon maka dari Tabel 6 dapat disimpulkan bahwa $H_0$ ditolak. Hal ini karena nilai *Likelihood Ratio Test* (LRT) sebesar 710,83 lebih besar dari $\chi^2_{0,05;9} = 16{,}919$. Artinya bahwa terdapat minimal satu variabel prediktor yang berpengaruh signifikan terhadap variabel respon. Berdasarkan pengujian secara parsial masing-masing variabel prediktor diperoleh

1. Pada model *count* dengan $\alpha = 0{,}05$ diperoleh nilai *p-value* < $\alpha = 0{,}05$ hanya variabel Log $X_3$ berarti bahwa variabel Log $X_3$ berpengaruh secara signifikan terhadap jumlah konsumsi rokok ($Y$). Sedangkan untuk variabel $X_1$ dan $X_2$ tidak berpengaruh secara signifikan terhadap jumlah konsumsi rokok ($Y$).
2. Pada model *zero inflation* dengan $\alpha = 0{,}05$ diperoleh semua nilai *p-value* < $\alpha = 0{,}05$ yang berarti bahwa semua variabel prediktor berpengaruh secara signifikan terhadap jumlah konsumsi rokok ($Y$).

Interpretasi model regresi ZINB adalah

1. Model *count* pada model regresi ZINB menjelaskan jumlah konsumsi rokok yang dipengaruhi oleh variabel prediktor yang signifikan. Untuk variabel Log penghasilan (Log $X_3$), setiap kenaikan satu Log penghasilan (Log $X_3$), maka jumlah konsumsi rokok setiap harinya akan bertambah sebanyak sebesar exp(0,1170437) = 1,124 batang rokok. Artinya semakin meningkat penghasilan seseorang, maka jumlah konsumsi rokok setiap harinya juga akan meningkat.
2. Model *Zero Inflation* pada model regresi ZINB menjelaskan peluang respon $y_i$ bernilai nol dipengaruhi oleh ketiga variabel prediktor yang berpengaruh signifikan terhadap jumlah konsumsi rokok ($Y$).
   a. Untuk variabel pendidikan ($X_1$), ketika pendidikan seseorang naik satu jenjang, maka peluang untuk tidak merokok adalah exp(0,47032) = 1,6 kali lebih mungkin dibanding merokok. Dengan kata lain, peluang seseorang yang memiliki jenjang pendidikan lebih tinggi untuk tidak merokok lebih besar daripada merokok.
   b. Untuk variabel kesejahteraan ($X_2$), ketika kesejahteraan seseorang naik satu jenjang, maka peluang untuk tidak merokok adalah exp(0,40274) = 1,496 kali lebih mungkin dibanding merokok. Dengan kata lain, peluang seseorang yang memiliki jenjang kesejahteraan lebih tinggi untuk tidak merokok lebih besar daripada merokok.
   c. Untuk variabel Log penghasilan (Log $X_3$), setiap kenaikan satu Log penghasilan (Log $X_3$), maka peluang untuk tidak merokok adalah exp(-0,35844) = 0,699 kali lebih mungkin dibanding merokok. Dengan kata lain, peluang seseorang yang memiliki penghasilan lebih tinggi untuk tidak merokok lebih kecil daripada merokok.

### 3.6. Model Regresi *Hurdle Negative Binomial* (HNB)

Model regresi *Hurdle Negative Binomial* (HNB) digunakan untuk menangani *overdispersi* pada model ZIP. Model regresi *Hurdle Negative Binomial* (HNB) terdiri dari model *count* atau model *truncated negative binomial* dan model *zero hurdle* atau model logit. Pada Tabel 7 disajikan estimasi dan pengujian parameter regresi HNB.

**Tabel 7.** Estimasi dan Pengujian Parameter Regresi HNB

| Variabel | Estimasi | Satndar Error | Nilai Z | *p-value* |
|---|---|---|---|---|
| Model *Count* atau Model *Truncated Negative Binomial* | | | | |
| Intercept | 0,8665214 | 0,1771272 | 4,892 | 9,98e-07 |
| $X_1$ | -0,0065090 | 0,0130030 | -0,501 | 0,617 |
| $X_2$ | 0,0009295 | 0,0136506 | 0,068 | 0,946 |
| Log $X_3$ | 0,1167654 | 0,0129265 | 9,033 | < 2e-16 |
| Log (theta) | 1,1128607 | 0,0362819 | 30,673 | < 2e-16 |
| Model *Zero Hurdle* atau Model Logit | | | | |
| Intercept | -2,67908 | 0,35433 | -7,561 | 4e-14 |
| $X_1$ | -0,46555 | 0,02915 | -15,969 | < 2e-16 |
| $X_2$ | -0,39948 | 0,03260 | -12,254 | < 2e-16 |
| Log $X_3$ | 0,36063 | 0,02657 | 13,571 | < 2e-16 |
| AIC = 26.656,63 | | Log *Likelihood* = -1,332e+04 | | |
| $\theta$ = 3,0431 | | LRT = 710,25 | | |

Pengujian parameter regresi HNB didasarkan hasil pada Tabel 7. Menggunakan $H_0$ adalah variabel prediktor tidak berpengaruh signifikan terhadap variabel respon maka dari Tabel 7 dapat disimpulkan bahwa $H_0$ ditolak. Hal ini karena nilai *Likelihood Ratio Test* (LRT) sebesar 710,25 lebih besar dari $\chi^2_{0,05;9} = 16{,}919$. Artinya bahwa minimal terdapat satu variabel prediktor yang

berpengaruh signifikan terhadap variabel respon. Berdasarkan pengujian secara parsial masing-masing variabel prediktor diperoleh

1. Pada model *count* dengan $\alpha = 0{,}05$ diperoleh nilai *p-value* < $\alpha = 0{,}05$ hanya variabel Log $X_3$ berarti bahwa variabel variabel Log $X_3$ berpengaruh secara signifikan terhadap jumlah konsumsi rokok ($Y$). Sedangkan untuk variabel $X_1$ dan $X_2$ tidak berpengaruh secara signifikan terhadap jumlah konsumsi rokok ($Y$).
2. Pada model *zero Hurdle* dengan $\alpha = 0{,}05$ diperoleh semua nilai *p-value* < $\alpha = 0{,}05$ yang berarti bahwa semua variabel prediktor berpengaruh secara signifikan terhadap jumlah konsumsi rokok ($Y$).

Interpretasi model regresi HNB adalah

1. Model *count* pada model regresi HNB menjelaskan jumlah konsumsi rokok yang dipengaruhi oleh variabel prediktor yang signifikan. Untuk variabel Log penghasilan (Log $X_3$), setiap kenaikan satu Log penghasilan (Log $X_3$), maka jumlah konsumsi rokok setiap harinya akan bertambah sebanyak sebesar exp(0,1167654) = 1,124 batang rokok. Artinya semakin meningkat penghasilan sesorang, maka jumlah konsumsi rokok setiap harinya juga akan meningkat.
2. Model *Zero Hurdle* pada model regresi HNB menjelaskan peluang respon $y_i$ bernilai nol dipengaruhi oleh ketiga variabel prediktor yang berpengaruh signifikan terhadap jumlah konsumsi rokok ($Y$).
   a. Untuk variabel pendidikan ($X_1$), ketika pendidikan seseorang naik satu jenjang, maka peluang untuk merokok adalah exp(-0,46555) = 0,628 kali lebih mungkin dibanding tidak merokok. Dengan kata lain, ketika pendidikan seseorang naik satu jenjang, maka peluang untuk tidak merokok adalah 1,593 kali lebih mungkin dibanding merokok.
   b. Untuk variabel kesejahteraan ($X_2$), ketika kesejahteraan seseorang naik satu jenjang, maka peluang untuk merokok adalah exp(-0,39948) = 0,671 kali lebih mungkin dibanding tidak merokok. Dengan kata lain, ketika kesejahteraan seseorang naik satu jenjang, maka peluang untuk tidak merokok adalah 1,491 kali lebih mungkin dibanding merokok.
   c. Untuk variabel Log penghasilan (Log $X_3$), setiap kenaikan satu Log penghasilan (Log $X_3$), maka peluang untuk merokok adalah exp(0,36063) = 1,434 kali lebih mungkin dibanding tidak merokok. Dengan kata lain, setiap kenaikan satu Log penghasilan (Log $X_3$), maka peluang untuk tidak merokok adalah 0,697 kali lebih mungkin dibanding merokok.

### 3.7. Pengecekan *Overdispersi*

Untuk menguji asumsi *overdispersi* pada regresi keluarga dilakukan uji dengan cara membagi nilai *Pearson Chi-Square* (Devian) dengan derajat bebasnya. Jika nilai tersebut lebih dari 1, maka dikatakan terjadi overdispersi. Perhitungan *Pearson Chi-Square* (Devian) dibagi dengan derajat bebasnya adalah

$$\hat{\phi} = \frac{63.178}{6.209} = 10{,}175$$

Karena nilai rasio dispersi $\hat{\phi} = 10{,}175 > 1$, maka dapat dikatakan bahwa pada regresi Poisson terjadi *overdispersi*.

### 3.8. Penentuan Model Terbaik

Pemilihan model terbaik yang digunakan dengan melihat nilai *Akaike's Information Criterion* (AIC) dari masing-masing model yang ditampilkan pada Tabel 8. Dilihat dari nilai AIC, dapat disimpulkan bahwa nilai AIC dari model regresi ZIP memiliki nilai AIC paling besar.

**Tabel 8**. Perbandingan Nilai AIC

| Model | AIC |
|---|---|
| Regresi ZIP | 32.919,80 |
| Regresi ZINB | 26.656,05 |
| Regresi HNB | 26.656,63 |

Nilai AIC model regresi ZINB dan model regresi HNB relatif sama tetapi nilai AIC model regresi ZINB lebih kecil dibandingkan nilai AIC model regresi HNB. Hal ini berarti baik model regresi ZINB dan model regresi HNB sudah mampu mengendalikan permasalahan *excess zero* dan *overdispersi*. Namun pada data jumlah konsumsi rokok model ZINB lebih baik digunakan untuk mengendalikan permasalahan *excess zero* dan *overdispersi* pada data tersebut.

## 4. KESIMPULAN

Berdasarkan kriteria nilai AIC dengan membandingkan model regresi ZIP, model regresi ZINB dan model regresi HNB dapat disimpulkan bahwa model regresi ZINB lebih baik digunakan untuk memodelkan data jumlah konsumsi rokok harian penduduk dewasa di Indonesia yang mengandung *excess zero* dan *overdispersi*. Model regresi ZINB yang diperoleh menunjukkan bahwa

1. Seseorang yang memiliki jenjang pendidikan lebih tinggi memiliki peluang lebih besar untuk tidak merokok..
2. Seseorang yang memiliki jenjang kesejahteraan lebih tinggi memiliki peluang lebih besar untuk tidak merokok.
3. Seseorang yang memiliki penghasilan lebih tinggi memiliki peluang lebih besar untuk merokok.

## DAFTAR PUSTAKA

[1]. Annisa, R., Permatasari, E.O., and Rumiati, A.T. 2020. Modeling of infant mortality in west sulawesi using zero inflated poisson regression method. *Journal of Physics: Conference Series* 1490.

[2]. Badan Pusat Statistik. 2020. *Statistik Indonesia 2020*. Jakarta : Badan Pusat Statistik.

[3]. Badan Pusat Statistik. 2020. *Statistik Kesejahteraan Rakyat 2020*. Jakarta : Badan Pusat Statistik.

[4]. Cameron, A.C. and Trivedi, P.K., 1998. *Regression Analysis of Count Data*. New York : Cambridge University Press.

[5]. Fang, R., Wagner, B.D., Harris, J.K., and Fillon, S.A. 2016. Zero-inflated negative binomial mixed model: an application to two microbial organisms important in oesophagitis. *Epidemiol Infect* 144 : 2447–2455.

[6]. Gujarati, D. 2009. *Dasar-Dasar Ekonometrika Jilid 2*. Jakarta: Erlangga.
[7]. Hilbe, J.M. 2011. *Negative Binomial Regression*, Second Edition. New York: Cambrigde University Press.
[8]. Hosmer, D.W. and Lemeshow, S. 2000. *Applied Logistic Regression*. New York: John Wiley & Sons, Inc.
[9]. Jansakul, N. and Hinde, J.P. 2002. Score Tests for Zero-Inflated Poisson Models. *Computational Statistics & Data Analysis.* Vol. 40 No.1 : 75-96.
[10]. Myers, R.H. 1990. *Classical and Modern Regression with Appilications*, 2nd ed. Boston: PW-KENT Publishing Company Boston.
[11]. Myers, R.H., Montgomery, D.C, Vining, G.G. and Robinson, T.J. 2010. *Generalized Linear Models with Applications in Engineering and The Sciences,* Second edition. New Jersey: John Wiley and Sons.
[12]. Nargis, N., Yong, H.H., Driezen, P., *et al*. 2019. Socioeconomic patterns of smoking cessation behavior in low and middle-income countries: Emerging evidence from the Global Adult Tobacco Surveys and International Tobacco Control Surveys. *PLoS ONE* 14(9): e0220223.
[13]. Osuji, G. A., Okoro, C. N., Obubu, M. and Obiora-Ilouno, H. O. 2016. Effect of Akaike Information Criterion on Model Selection in Analyzing Auto-crash Variables. *International Journal of Sciences: Basic and Applied Research (IJSBAR)*. Volume 26 No 1 : 98-109.
[14]. Wang, Q., Shen, J.J., Sotero, M., *et al*. 2018. Income, occupation and education : Are they related to smoking behaviors in China?. *PLoS ONE* 13(2): e0192571.
[15]. Zhen, Z., Shao, L. and Zhang, L. 2018. Spatial Hurdle Models for Predicting the Number of Children with Lead Poisoning. *International Journal of Environmental Research and Public Health* 15.