Pemodelan Harga Mobil Bekas Menggunakan *Generalized Linear Model*

TUGAS BESAR

Sebagai salah satu penilaian
Kuliah AK4082 Model Linier Lanjut

Oleh

10820033 Pamella Cathryn

Program Studi Aktuaria
Fakultas Matematika dan Ilmu Pengetahuan Alam
Institut Teknologi Bandung

2023

Daftar Isi

Halaman

Daftar Tabel

Daftar Gambar

1 Pendahuluan

1.1 Latar Belakang

Pasar mobil bekas adalah pasar yang cukup kompleks dan dinamis, dimana terdapat berbagai faktor yang mempengaruhi harga mobil bekas tersebut. Beberapa faktor diantaranya adalah kilometer tempuh mobil, transmisi gigi mobil, dan lain-lain. Karena banyaknya variabel yang terlibat, pemodelan harga mobil bekas dapat menjadi sesuatu yang menantang. Pengembangan model harga mobil bekas yang akurat sangat penting baik untuk pembeli maupun penjual. Hal ini karena dapat memberikan informasi untuk pengambilan keputusan harga yang tepat dan memastikan nilai pasar yang adil.

Dalam karya tulis ini, akan digunakan metode *Generalized Linear Model* untuk memodelkan harga mobil bekas. GLM menawarkan pendekatan yang fleksibel untuk menganalisis data yang kompleks dan telah digunakan secara luas di berbagai industri untuk bermacam-macam aplikasi. Dengan menerapkan metode ini, akan diperoleh sebuah model prediksi harga mobil bekas yang akurat berdasarkan variabel-variabel prediktornya.

1.2 Metodologi

Pada karya tulis ini, akan dilakukan pemodelan harga mobil bekas menggunakan metode *generalized linear model* (GLM). Metode tersebut digunakan karena harga mobil bekas sebagai variabel respons tidak mengikuti distribusi normal. Harga mobil bekas sendiri merupakan data kontinu nonnegatif dan *skewed to the right*. Sehingga, distribusi variabel respons nantinya akan diasumsikan berasal dari distribusi gamma dan inverse gaussian.

1.2.1 *Generalized Linear Model* (GLM)

Pada metode GLM, variabel respons diasumsikan mengikuti suatu distribusi dari keluarga distribusi eksponensial dimana fungsi peluangnya dapat ditulis dalam bentuk:

$$f(y) = c(y, \phi) \exp\left(\frac{y\theta - a(\theta)}{\phi}\right) \tag{1}$$

yang mana, y adalah respons, ϕ adalah parameter dispersi, dan θ adalah parameter kanonik.

Pada metode GLM, fungsi rataan respons dimodelkan secara linier dengan variabel-variabel prediktornya. Sehingga, persamaan model yang diperoleh dengan metode GLM adalah:

$$g(\mu) = \mathbf{x}'\beta \tag{2}$$

yang mana, $g(.)$ merupakan *link function*, $\mathbf{x}$ merupakan vektor dari variabel-variabel prediktor, dan β merupakan vektor dari parameter-parameter yang bersesuaian dengan variabel prediktor.

1.2.2 Asumsi Distribusi Variabel Respons

Data yang akan digunakan sebagai variabel respons pada karya tulis ini adalah data harga mobil bekas yang merupakan data kontinu nonnegatif. Sehingga, variabel respons akan diasumsikan berasal dari distribusi gamma dan distribusi inverse gaussian.

1. Distribusi Gamma

   Fungsi peluang dari $Y \sim G(\mu, \nu)$ adalah:

$$f(y) = \frac{y^{-1}}{\Gamma(\nu)} \left(\frac{y\nu}{\mu}\right)^{\nu} e^{-\frac{y\nu}{\mu}}, \quad y > 0 \tag{3}$$

   Distribusi gamma berasal dari keluarga distribusi eksponensial karena persaamaan (3) dapat ditulis ulang ke bentuk persamaan (1) dengan $\theta = -\frac{1}{\mu}$, $a(\theta) = -\ln(-\theta)$, dan $\phi = \frac{1}{\nu}$. *Canonical link* dari distribusi gamma adalah $g(\mu) = \mu^{-1}$.

2. Distribusi Inverse Gaussian

   Fungsi peluang dari $Y \sim IG(\mu, \sigma^2)$ adalah:

$$f(y) = \frac{1}{\sqrt{2\pi y^3 \sigma^2}} \exp\left(-\frac{1}{2y}\left(\frac{y-\mu}{\mu\sigma}\right)^2\right), \quad y > 0 \tag{4}$$

   Distribusi inverse gaussian berasal dari keluarga distribusi eksponensial karena persaamaan (4) dapat ditulis ulang ke bentuk persamaan (1) dengan $\theta = -\frac{1}{2\mu^2}$, $a(\theta) = -\sqrt{-2\theta}$, dan $\phi = \sigma^2$. *Canonical link* dari distribusi inverse gaussian adalah $g(\mu) = \mu^{-2}$.

1.2.3 *Maximum Likelihood Estimation* (MLE)

Fungsi *log-likelihood* dari persamaan 1 adalah:

$$\begin{aligned} \ell(\beta, \phi) &= \sum_{i=1}^{n} \ln f(y_i; \beta, \phi) \\ &= \sum_{i=1}^{n} \left(\ln c(y_i, \phi) + \frac{y_i\theta_i - a(\theta_i)}{\phi}\right) \end{aligned} \tag{5}$$

Estimasi nilai β dan ϕ dapat diperoleh dari menurunkan secara parsial persamaan 5 terhadap masing-masing parameter. Sehingga, akan diperoleh suatu matriks:

$$\mathbf{X}'D(y-\mu) = 0 \Leftrightarrow \mathbf{X}'WG(y-\mu) = 0 \tag{6}$$

yang mana $\mathbf{X}$ adalah *design matrix* dengan ukuran $n \times (m+1)$ dengan n adalah banyak observasi dan m adalah banyak kolom dan D adalah matriks diagonal dengan entri $\partial\theta_i/\partial\eta_i$ yang mana $\eta_i = x_i\beta$.

1.2.4 *Deviance*

Untuk melihat apakah model yang diperoleh cocok untuk memodelkan data, maka akan digunakan *deviance*. *Deviance* adalah ukuran jarak antara *saturated* model ($\check{\ell}$) dengan

model yang diperoleh ($\hat{\ell}$). Berikut merupakan persamaannya:

$$\Delta \equiv 2(\,\check{\ell} - \hat{\ell}\,) = 2\sum_{i=1}^{n}\left(\frac{y_i(\check{\theta}_i - \hat{\theta}_i) - a(\check{\theta}_i) + a(\hat{\theta}_i)}{\phi}\right) \tag{7}$$

Ketika model cocok dengan data, maka nilai $\hat{\ell}$ akan dekat (tapi tidak lebih besar) dengan $\check{\ell}$. Residu pada GLM berbeda dengan residu pada model regresi linier klasik karena tidak berdistribusi normal dan tidak memiliki variansi yang konstan. Pada karya tulis ini, residu model akan ditinjau dengan *deviance residuals* (dinotasikan dengan δ_i^2) yang memiliki persamaan yang sama dengan persamaan (7), yaitu:

$$\delta_i^2 \equiv 2\left(\frac{y_i(\check{\theta}_i - \hat{\theta}_i) - a(\check{\theta}_i) + a(\hat{\theta}_i)}{\phi}\right) \tag{8}$$

Jika diperoleh nilai *deviance residuals* $|\delta_i| > 1$, maka dapat dikatakan bahwa model yang diperoleh kurang cocok dengan data.

2 Analisis Data

Data yang akan digunakan pada karya tulis ini adalah data informasi mobil bekas yang dijual pada CarDekho.com (sebuat *platform* jual-beli otomotif di India) yang diambil melalui metode *web scraping* pada tahun 2020. Data ini diperoleh dari kaggle.com pada April 2023.

2.1 Pemahaman dan Pembersihan Data (*Data Understanding and Cleaning*)

Data kotor memiliki 4340 baris dan 8 kolom (1 kolom variabel respons dan 7 kolom variabel prediktor) dengan tidak ada *missing values* pada data. Data ini akan dipecah menjadi dua, yaitu data *train* sebanyak 3472 (80%) dan data *test* sebanyak 868 (20%). Berikut merupakan cuplikan 5 baris pertama dari data kotor tersebut:

Tabel 2.1. 5 Baris Pertama Data Kotor

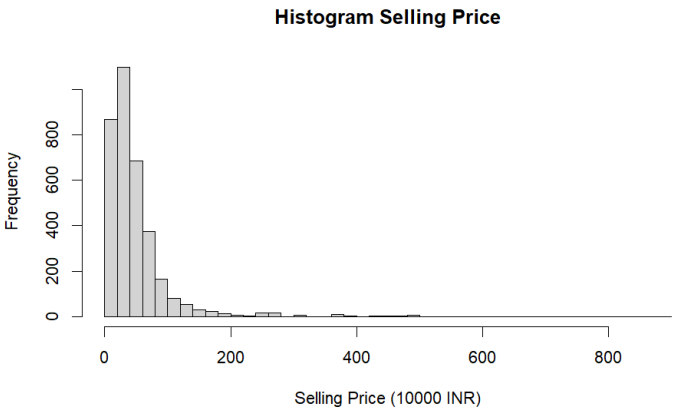
Selling Price	Name	Year	Km Driven	Fuel	Seller Type	Transmission	Owner
60000	Maruti 800 AC	2007	70000	Petrol	Individual	Manual	First Owner
135000	Maruti Wagon R LXI Minor	2007	50000	Petrol	Individual	Manual	First Owner
600000	Hyundai Verna 1.6 SX	2012	100000	Diesel	Individual	Manual	First Owner
250000	Datsun RediGO T Option	2017	46000	Petrol	Individual	Manual	First Owner
450000	Honda Amaze VX i-DTEC	2014	141000	Diesel	Individual	Manual	Second Owner

2.1.1 *Selling Price*

Kolom ini merupakan kolom data kontinu harga jual mobil bekas dalam satuan Indian Rupee (INR). Kolom inilah yang akan menjadi variabel respons pada pemodelan ini. Untuk mengefisienkan pemodelan, kolom ini akan di-*convert* ke dalam satuan 10000 INR. Berikut adalah statistika deskriptifnya:

Tabel 2.2. Statistika Deskriptif *Selling Price*

Statistik	Nilai
Minimum	2.00
Maksimum	890.00
Mean	50.08
Kuartil 1	20.07
Median	35.00
Kuartil 3	60.00



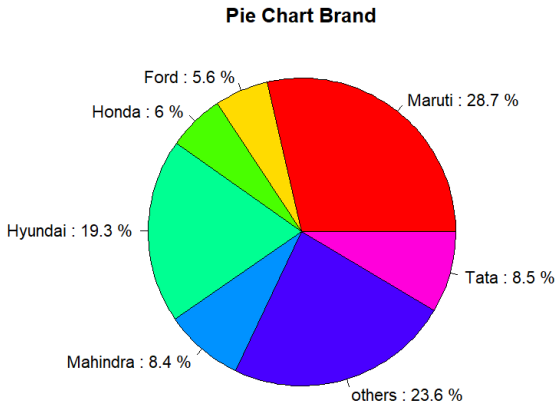
Gambar 2.1. Histogram *Selling Price*

2.1.2 *Name*

Kolom *Name* ini merupakan kolom yang berisi nama unit mobil bekas yang dijual. Dari kolom ini, akan diambil merek mobil bekas tersebut dan di simpan ke sebuah kolom data kategorikal bernama *Brand* yang akan menggantikan kolom *Name*. Merek-merek mobil bekas yang kurang banyak dijual akan disimpan dengan nilai *"others"*. Sehingga, kolom *Brand* ini berisi 7 merek mobil bekas yang dijual, yaitu: Maruti, Hyundai, Mahindra, Tata, Honda, Ford, dan *others*. Perhatikan pada gambar 2.2, merek Maruti memiliki frekuensi yang paling besar. Sehingga, *base level* dari variabel ini adalah merek Maruti. Berikut adalah statistika deskriptifnya:

Tabel 2.3. Statistika Deskriptif *Brand*

Kategori	Frekuensi
Maruti	995
Hyundai	671
Mahindra	290
Tata	296
Honda	208
Ford	194
others	818



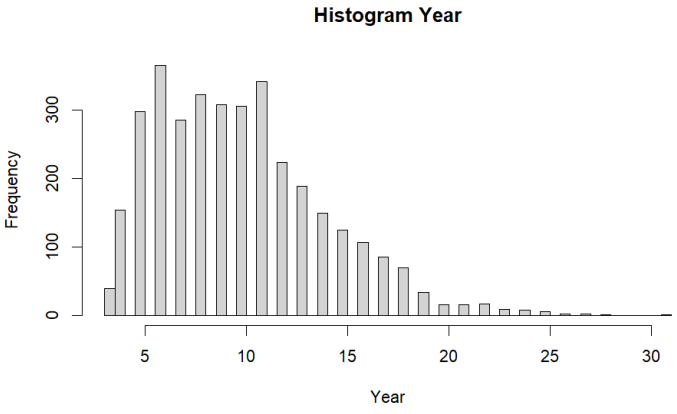
Gambar 2.2. Diagram Lingkaran *Brand*

2.1.3 *Year*

Kolom ini berisi data diskret tahun saat mobil bekas yang dijual pertama kali dibeli. Kolom ini akan diganti menjadi 2023-*Year* agar lebih numerik dan objektif. Kolom ini pun akan ditinjau sebagai salah satu dari *offset* pada pemodelan ini. Berikut merupakan statistika deskriptifnya:

Tabel 2.4. Statistika Deskriptif *Year*

Statistik	Nilai
Minimum	3
Maksimum	31
Mean	9.906
Kuartil 1	7
Median	9
Kuartil 3	12



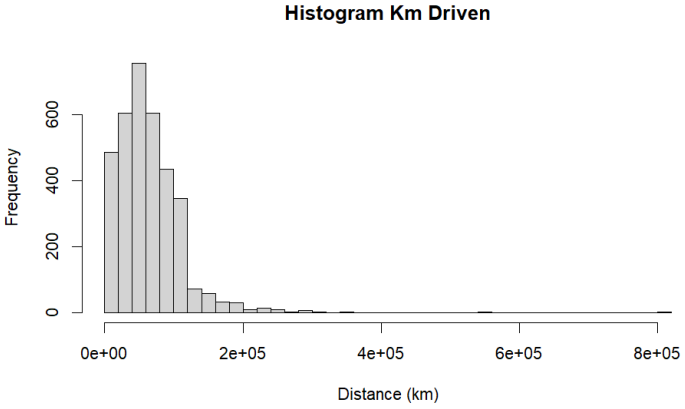
Gambar 2.3. Histogram *Year*

2.1.4 Km *Driven*

Kolom ini berisi total jarak yang telah ditempuh oleh mobil bekas dalam satuan kilometer. Data jarak ini merupakan data kontinu nonnegatif dan akan ditinjau sebagai salah satu dari *offset* pada pemodelan ini. Berikut adalah statistika deskriptifnya:

Tabel 2.5. Statistika Deskriptif *Km Driven*

Statistik	Nilai
Minimum	1
Maksimum	806599
Mean	66828
Kuartil 1	35000
Median	60000
Kuartil 3	90000



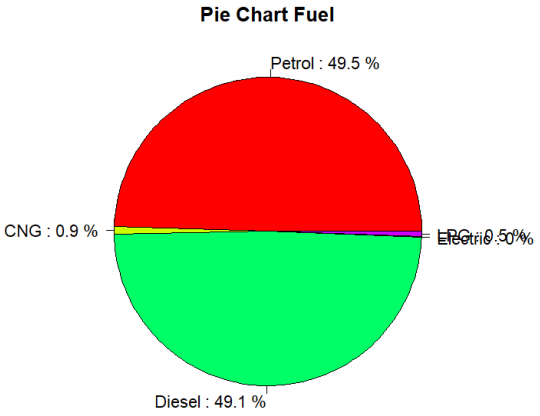
Gambar 2.4. Histogram *Km Driven*

2.1.5 *Fuel*

Kolom ini berisi data jenis bahan bakar dari mobil bekas yang dijual. Data ini merupakan data kategorikal yang berisi kategori-kategori: Petrol, Diesel, CNG, LPG, Electric. Bahan bakar jenis Petrol akan menjadi *base level* karena nilai frekuensinya yang paling besar menurut 2.5. Berikut adalah statistika deskriptifnya:

Tabel 2.6. Statistika Deskriptif *Fuel*

Kategori	Frekuensi
Petrol	1717
Diesel	1705
CNG	30
LPG	19
Electric	1



Gambar 2.5. Diagram Lingkaran *Fuel*

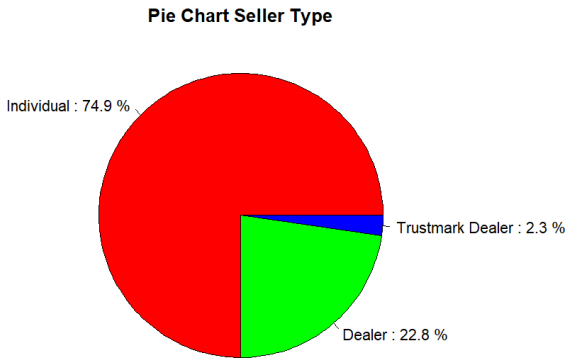
2.1.6 *Seller Type*

Kolom ini berisi data kategorikal mengenai jenis penjual mobil bekas. Jenis yang dimaksud adalah individu, *dealer*, atau *trustmark dealer*. Pada gambar 2.6, tipe penjual individu

memiliki frekuensi yang paling besar. Sehingga, tipe penjual individu akan menjadi *base level*. Berikut adalah statistika deskriptifnya:

Tabel 2.7. Statistika Deskriptif *Seller Type*

Kategori	Frekuensi
Individual	2602
Dealer	790
Trustmark Dealer	80



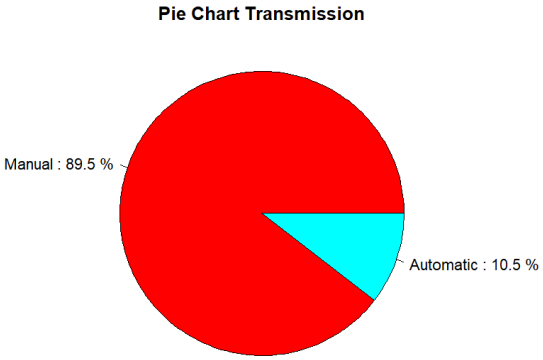
Gambar 2.6. Diagram Lingkaran *Seller Type*

2.1.7 *Transmission*

Kolom ini berisi data kategorikal jenis transmisi mobil bekas yang dijual, yaitu manual atau *automatic*. Jenis transmisi manual akan dijadikan *base level* karena memiliki frekuensi yang paling besar menurut gambar 2.8. Berikut adalah statistika deskriptifnya:

Tabel 2.8. Statistika Deskriptif *Transmission*

Kategori	Frekuensi
Manual	3108
Automatic	364



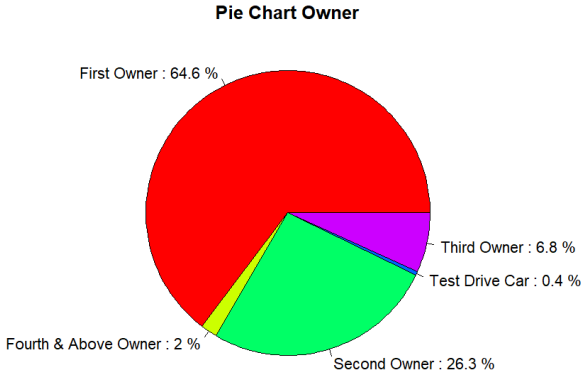
Gambar 2.7. Diagram Lingkaran *Transmission*

2.1.8 *Owner*

Kolom ini berisi data kategorikal mengenai jumlah pemilik mobil bekas sebelumnya. Kategori-kategori yang terdapat pada kolom ini adalah: *first owner*, *second owner*, *third owner*, *fourth & above owner*, dan *test drive car*. *First owner* akan menjadi *base level* karena memiliki frekuensi paling besar menurut gambar 2.8. Berikut adalah statistika deskriptifnya:

Tabel 2.9. Statistika Deskriptif *Owner*

Kategori	Frekuensi
First Owner	2243
Second Owner	912
Third Owner	235
Fourth & Above Owner	68
Test Drive Car	14



Gambar 2.8. Diagram Lingkaran *Owner*

Setelah data kotor dibersihkan, diperoleh data yang sudah bersih dan siap digunakan untuk dianalisis. Berikut adalah cuplikan 5 baris pertama dari data bersih:

Tabel 2.10. 5 Baris Pertama Data Bersih

Selling Price	Brand	Year	Km Driven	Fuel	Seller Type	Transmission	Owner
6	Maruti	16	70000	Petrol	Individual	Manual	First Owner
13.5	Maruti	16	50000	Petrol	Individual	Manual	First Owner
60	Hyundai	11	100000	Diesel	Individual	Manual	First Owner
25	others	6	46000	Petrol	Individual	Manual	First Owner
45	Honda	9	141000	Diesel	Individual	Manual	Second Owner

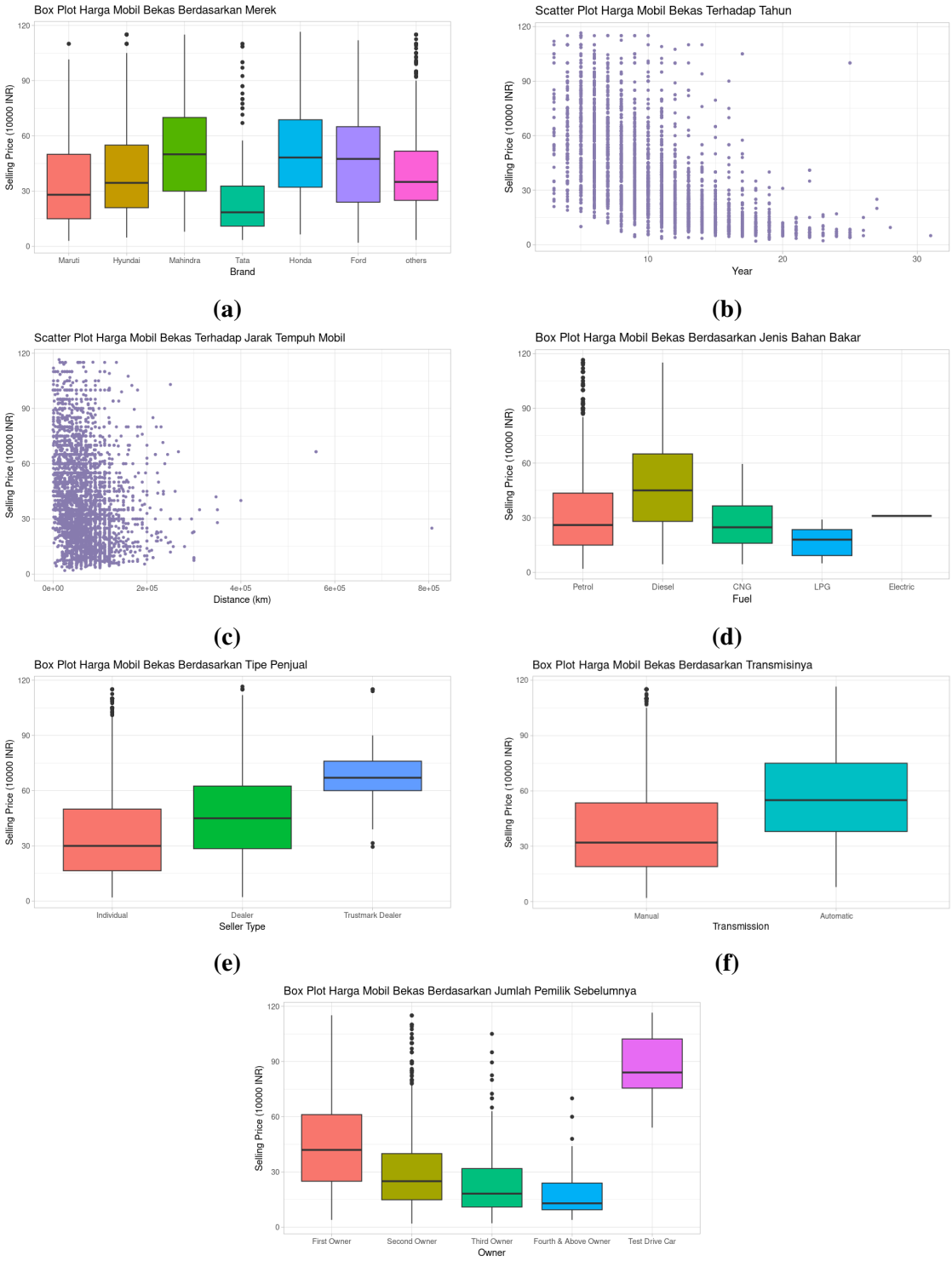
2.2 *Exploratory Data Analysis* (EDA)

Pada subbab ini, akan dilakukan eksplorasi terhadap hubungan antara variabel respons dengan variabel prediktor dan interaksi antarvariabel prediktor.

2.2.1 Hubungan Variabel Respons Terhadap Variabel Prediktor

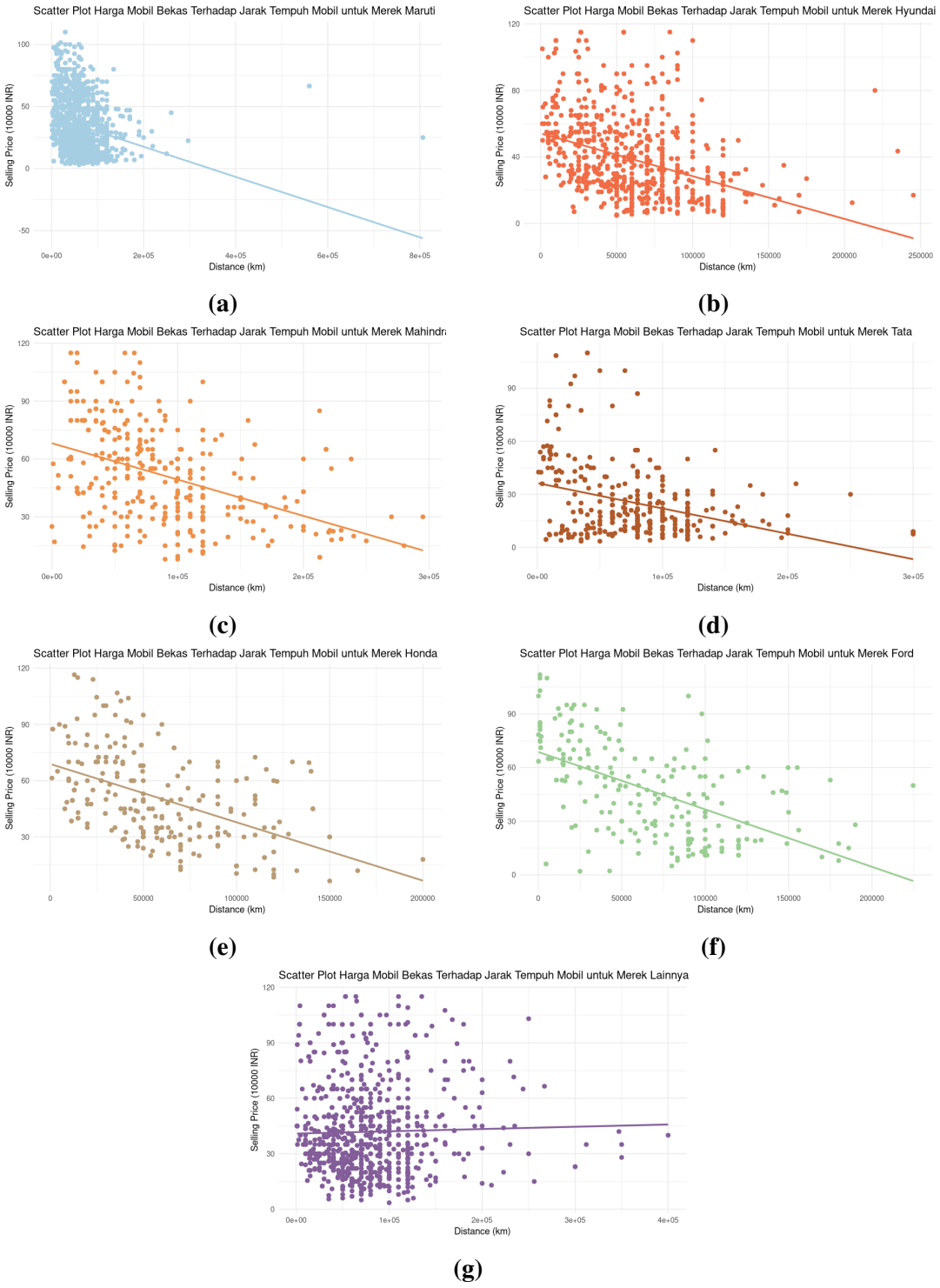
Gambar 2.9 menunjukkan beberapa visualisasi hubungan variabel respons (harga mobil bekas) terhadap variabel-variabel prediktornya. Pada gambar 2.9a, terlihat secara kasar bahwa persebaran harga mobil bekas tiap merek berada pada interval yang cenderung sama, kecuali untuk merek Tata yang tersebar pada harga yang lebih kecil dari merek lain. Pada gambar 2.9b, terlihat bahwa harga mobil bekas cenderung memiliki korelasi positif terhadap tahun pertama kali mobil dibeli. Dari gambar 2.9c, sedikit sulit untuk melihat hubungannya, sehingga akan dilakukan pengelompokkan berdasarkan merek yang dirincikan pada gambar 2.10. Pada gambar 2.9d, terlihat bahwa persebaran harga mobil bekas tertinggi dimiliki oleh kategori bahan bakar diesel, diikuti oleh petrol, CNG, lalu LPG. Untuk mobil elektrik, hanya terdapat 1 datum saja. Pada gambar 2.9e, terlihat bahwa persebaran harga mobil bekas tertinggi dimiliki oleh kategori tipe penjual *trustmark dealer*, diikuti oleh *dealer*, dan persebaran harga terendah oleh tipe penjual individu. Pada gambar 2.9f, terlihat bahwa mobil dengan transmisi manual cenderung lebih murah daripada mobil dengan transmisi *automatic*. Terakhir, pada gambar 2.9g, terlihat bahwa mobil *test drive*

memiliki persebaran harga yang paling tinggi, diikuti oleh mobil *first owner*, *second owner*, *third owner*, dan *fourth & above owner*.



Gambar 2.9. Hubungan Variabel Respons dengan Variabel Prediktor

Gambar 2.10 menunjukkan visualisasi hubungan antara harga mobil bekas terhadap jarak tempuh dengan pengelompokkan berdasarkan mereknya. Perhatikan bahwa harga mobil bekas cenderung memiliki hubungan berbanding terbalik terhadap jarak tempuhnya untuk seluruh merek selain merek lainnya (*others*).



Gambar 2.10. Hubungan Harga Mobil Bekas Terhadap Jarak Tempuh Dikelompokkan Berdasarkan Merek

2.2.2 Pengecekan Multikolinearitas

Selanjutnya, akan dicek multikolinearitas antarvariabel prediktor nonkategorikal, yaitu variabel *year* dan variabel km *driven*. Hasil berupa tabel korelasi dapat dilihat pada 2.11. Perhatikan bahwa variabel *year* dan variabel km *driven* memiliki hubungan linier yang positif namun nilai korelasi yang tidak besar. Sehingga, kedua variabel ini akan tetap digunakan pada pemodelan.

Tabel 2.11. Tabel Korelasi Variabel Prediktor Nonkategorikal

	Year	Km Driven
Year	1.00	0.42
Km Driven	0.42	1.00

2.2.3 Interaksi Antarvariabel Prediktor

Penentuan interaksi antarvariabel prediktor yang dimasukkan ke dalam pemodelan akan dilakukan menggunakan *prior knowledge*. Selain itu, diasumsikan bahwa interaksi yang mungkin terjadi terbatas pada antara dua variabel saja. Berikut ini merupakan beberapa interaksi antarvariabel prediktor yang akan ditinjau pada karya tulis ini, yaitu:

a. *Year* dan Km *Driven*

b. *Year* dan *Transmission*

c. *Year* dan *Owner*

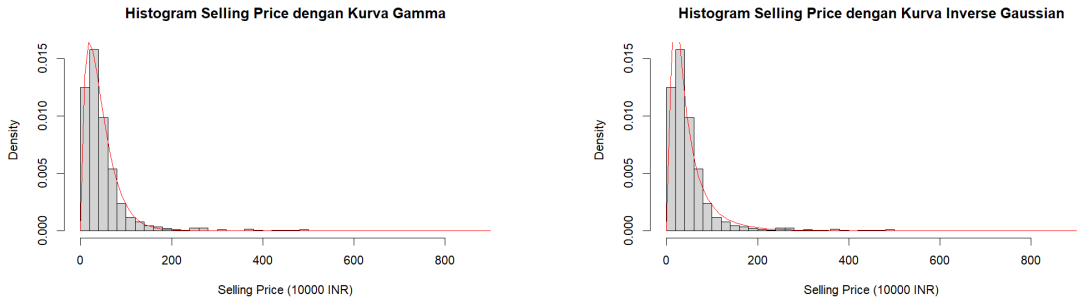
d. Km *Driven* dan *Owner*

2.3 Pengolahan Data (*Data Processing*)

Pada subbab ini, akan cari distribusi dari variabel respons. Berdasarkan asumsi distribusi respons pada 1.2.2, distribusi yang akan dicoba adalah distribusi gamma dan distribusi inverse gaussian. Parameter-parameternya akan diestimasi menggunakan metode *maximum likelihood estimation* (MLE) dengan bantuan R. Setelah itu, akan dilakukan uji kecocokan data dengan distribusi menggunakan uji Kolmogorov-Smirnov. Hasil yang diperoleh adalah sebagai berikut:

Tabel 2.12. Perbandingan Hasil Uji Kolmogorov-Smirnov Asumsi Distribusi Variabel Respons

Distribusi	Estimasi Parameter	p-value Uji Kolmogorov-Smirnov	AIC	BIC
Gamma $G(\mu, \nu)$	$\hat{\mu} = \hat{\alpha}/\hat{\beta} = 44.1193$ $\hat{\nu} = \hat{\alpha} = 1.8226$	**0.1201**	33971.99	33984.29
Inverse Gaussian $IG(\mu, \sigma^2)$	$\hat{\mu} = 48.6347$ $\hat{\sigma}^2 = \hat{\mu}^3/\hat{\lambda} = 1954.2706$	0.0803	**33390.71**	**33403.02**



(a) Histogram Selling Price dengan Kurva Gamma

(b) Histogram Selling Price dengan Kurva Inverse Gaussian

Gambar 2.11. Histogram Selling Price dengan Kurva Distribusi

Perhatikan bahwa, dengan taraf signifikansi 5%, nilai *p-value* hasil uji Kolmogorov-Smirnov kedua distribusi sama-sama menyimpulkan bahwa variabel respons mengikuti kedua distribusi (*p-value* > 0.05). Nilai AIC dan BIC yang dimiliki oleh distribusi inverse gaussian lebih kecil daripada nilai AIC dan BIC dari distribusi gamma meskipun nilai perbedaan AIC dan BIC-nya tidak jauh berbeda. Sehingga, kedua distribusi ini akan ditinjau sebagai asumsi distribusi respons pada pembangunan model.

2.4 Pemodelan Data (*Data Modeling*)

Pada subbab ini, akan dilakukan pemodelan dengan beberapa model sebagai berikut:

1. *full model* linier dengan asumsi distribusi gamma menggunakan link $g(\mu) = \mu^{-1}$
2. *full model* linier dengan asumsi distribusi inverse gaussian menggunakan link $g(\mu) = \ln \mu$
3. *full model* linier dengan asumsi distribusi gamma menggunakan link $g(\mu) = \ln \mu$
4. *full model* linier ditambah interaksi dengan asumsi distribusi gamma menggunakan link $g(\mu) = \mu^{-1}$
5. *full model* linier ditambah interaksi dengan asumsi distribusi inverse gaussian menggunakan link $g(\mu) = \ln \mu$
6. *full model* linier dimana variabel *year* dan variabel km *driven* sebagai *offset* dengan asumsi distribusi gamma menggunakan link $g(\mu) = \ln \mu$
7. *full model* linier dimana variabel *year* dan variabel km *driven* sebagai *offset* dengan asumsi distribusi inverse gaussian menggunakan link $g(\mu) = \ln \mu$

Untuk mempersingkat penulisan, akan didefinisikan variabel-variabel berikut:
y = variabel *selling price* (variabel respons)
x_1 = variabel *brand*
x_2 = variabel *year*
x_3 = variabel km *driven*

x_4 = variabel *fuel*
x_5 = variabel *seller type*
x_6 = variabel *transmission*
x_7 = variabel *owner*

Dengan bantuan R, diperoleh ringkasan untuk setiap model sebagai berikut:

Tabel 2.13. Ringkasan Hasil Pemodelan Calon Model

Model	Parameter Tak Signifikan	AIC	Null Deviance	Residual Deviance	MAPE
1	x_4CNG x_4Electric x_7Second Owner x_7Test Drive Car	30047	2521.56	908.71	0.515
2	x_1Ford x_4CNG x_4Electric x_7Test Drive Car	29641	**70.829**	**24.246**	0.389
3	x_4CNG x_4LPG x_4Electric x_7Test Drive Car	**29029**	2521.56	685.17	**0.374**
4	x_3 x_4CNG x_4LPG x_4Electric x_7Second Owner x_7Third Owner x_7Fourth & Above Owner x_7Test Drive Car x_2*x_3 x_2*x_7Test Drive Car x_3*x_7Test Drive Car	29334	2521.56	741.64	0.446
5	x_1Ford x_3 x_4CNG x_4Electric x_7Test Drive Car x_2*x_7Test Drive Car x_2*x_7Second Owner x_2*x_7Third Owner x_2*x_7Fourth & Above Owner x_3*x_7Test Drive Car	29556	1873.92	768.03	0.955

6	x_1Hyundai x_1Tata x_1Honda x_1Ford $x_1$others x_4Diesel x_4CNG x_4LPG x_4Electric x_5Dealer x_5Trustmar Dealer x_7Fourth & Above Owner	40418	20835	11703	>1
7	x_1Mahindra x_1Honda x_1Ford x_7Fourth & Above Owner	35263	140.16	122.55	>1

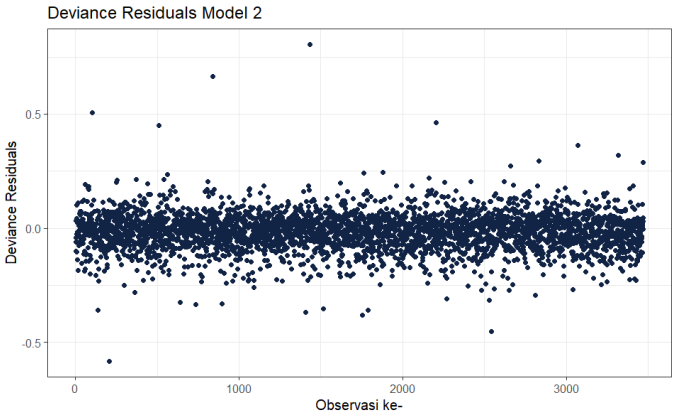
Dari tabel 2.13, diperoleh model dengan nilai AIC terendah adalah model 3, model dengan nilai *null deviance* terendah adalah model 2, model dengan nilai *residual deviance* terendah adalah model 2, dan model dengan nilai MAPE terendah adalah model 3. Sehingga, model yang akan dianalisis lebih lanjut adalah model 2 dan model 3.

2.5 Analisis Residu

Selanjutnya, akan dilihat performa model 2 dan model 3 berdasarkan *deviance residuals*-nya. Berikut merupakan informasi *deviance residuals* dari kedua model:

Tabel 2.14. Statistika Deskriptif *Deviance Residuals* Model 2

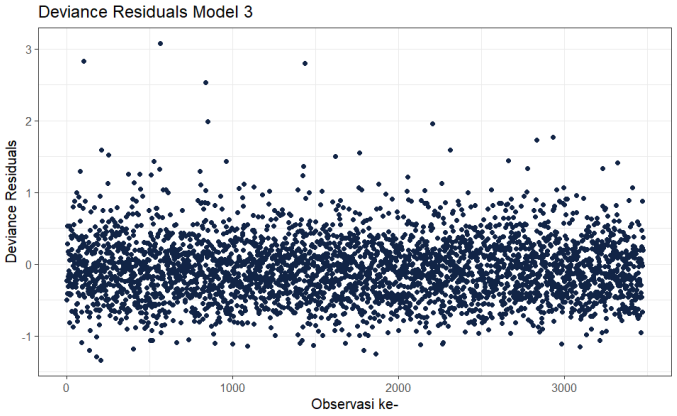
Statistik	Nilai
Minimum	-0.58399
Maksimum	0.80529
Mean	-0.01660
Kuartil 1	-0.06428
Median	-0.01285
Kuartil 3	0.02975



Gambar 2.12. Plot *Deviance Residuals* Model 2

Tabel 2.15. Statistika Deskriptif *Deviance Residuals* Model 3

Statistik	Nilai
Minimum	-1.33900
Maksimum	3.07654
Mean	-0.06574
Kuartil 1	-0.37457
Median	-0.08351
Kuartil 3	0.19824



Gambar 2.13. Plot *Deviance Residuals* Model 3

Suatu model GLM dapat dikatakan kurang baik apabila memiliki nilai *deviance residuals* (δ_i^2) yang cenderung jauh dari 0. Perhatikan bahwa nilai *deviance residuals* dari model 2 semuanya terletak pada interval $|\delta_i^2| < 1$ sementara model 3 tidak. Sehingga, dapat ditarik kesimpulan bahwa performa model 2 lebih baik daripada model 3.

2.6 Validasi Data *Test*

Selanjutnya, model 2 akan digunakan untuk memprediksi harga mobil bekas dengan menggunakan data *test*. Dengan bantuan R, diperoleh MAPE sebesar 0.4163 atau sebesar 41.63%. Sementara jika model 3 yang digunakan untuk memprediksi harga mobil bekas dengan menggunakan data *test*, akan diperoleh MAPE sebesar 0.5518 atau sebesar 55.18%. Hal ini mendukung argumen sebelumnya bahwa performa model 2 lebih baik daripada model 3.

3 Kesimpulan dan Saran

3.1 Model yang Diperoleh

Dari Bab 2, diperoleh model yang terbaik untuk pemodelan harga mobil bekas adalah model 2, yaitu *full model* linier dengan asumsi distribusi inverse gaussian menggunakan link $g(\mu) = \ln \mu$. Dengan bantuan R, diperoleh persamaan model sebagai berikut:

$$\begin{aligned}\ln \mu = {} & 4.243 + 0.07442\, x_1\text{Hyundai} + 0.2811\, x_1\text{Mahindra} - 0.3416\, x_1\text{Tata} \\ & + 0.4085\, x_1\text{Honda} + 0.03714\, x_1\text{Ford} + 0.2226\, x_1\text{others} - 0.09804\, x_2 \\ & - 5.482 \times 10^{-7}\, x_3 + 0.4685\, x_4\text{Diesel} - 0.08021\, x_4\text{CNG} - 0.1808\, x_4\text{LPG} \\ & - 0.2075\, x_4\text{Electric} + 0.06682\, x_5\text{Dealer} + 0.5092\, x_5\text{Trustmark Dealer} \\ & + 0.8702\, x_6\text{Automatic} - 0.06067\, x_7\text{Second Owner} - 0.1631\, x_7\text{Third Owner} \\ & - 0.1322\, x_7\text{Fourth \& Above Owner} + 0.2984\, x_7\text{Test Drive Car}\end{aligned}$$

yang mana,
μ = *expected value* dari variabel response
x_1Hyundai = variabel *brand* level Hyundai = $\{0, 1\}$
x_1Mahindra = variabel *brand* level Mahindra = $\{0, 1\}$
x_1Tata = variabel *brand* level Tata = $\{0, 1\}$
x_1Honda = variabel *brand* level Honda = $\{0, 1\}$
x_1Ford = variabel *brand* level Ford = $\{0, 1\}$
$x_1$others = variabel *brand* level others = $\{0, 1\}$
x_2 = variabel *year*
x_3 = variabel km *driven*
x_4Diesel = variabel *fuel* level Diesel = $\{0, 1\}$
x_4CNG = variabel *fuel* level CNG = $\{0, 1\}$
x_4LPG = variabel *fuel* level LPG = $\{0, 1\}$
x_4Electric = variabel *fuel* level Electric = $\{0, 1\}$
x_5Dealer = variabel *seller type* level Dealer = $\{0, 1\}$
x_5Trustmark Dealer = variabel *seller type* level Trustmark Dealer = $\{0, 1\}$
x_6Automatic = variabel *transmission* level Automatic = $\{0, 1\}$
x_7Second Owner = variabel *owner* level Second Owner = $\{0, 1\}$
x_7Third Owner = variabel *owner* level Third Owner = $\{0, 1\}$
x_7Fourth & Above Owner = variabel *owner* level Fourth & Above Owner = $\{0, 1\}$
x_7Test Drive Car = variabel *owner* level Test Drive Car = $\{0, 1\}$

Sebagai tambahan, faktor-faktor yang memengaruhi harga mobil bekas dapat bermacam-macam. Pada dua model terbaik dari seluruh calon model, dapat disimpulkan bahwa:

- Jika parameter yang digunakan adalah *null deviance* dan *residual deviance*, maka variabel prediktor x_1Ford, x_4CNG, x_4Electric, dan x_7Test Drive Car tidak terlalu berkontribusi dalam penentuan harga mobil bekas.
- Jika parameter yang digunakan adalah AIC dan MAPE, maka variabel prediktor x_4CNG, x_4LPG, x_4Electric, dan x_7Test Drive Car tidak terlalu berkontribusi dalam penentuan harga mobil bekas.

3.2 Saran

Dapat dilihat bahwa MAPE model yang ditinjau tidak ada yang dibawah 10%. Hal ini dapat mengindikasikan bahwa model yang diperoleh masih kurang cocok pada data atau data yang digunakan belum cukup dalam memodelkan harga mobil bekas. Sebagai saran peningkatan model, sebaiknya ditambah variabel-variabel prediktor, meninjau data pencilan, dan perbanyak data.

3.3 Contoh Kasus

Misalkan, ada seseorang yang hendak menjual mobil bekas *first hand*-nya. Mobil tersebut bermerek Tata, dibeli pada tahun 2013, memiliki jarak tempuh 50000 km, berbahan bakar Petrol, dan memiliki transmisi manual. Maka:

$$\begin{aligned}\ln \mu = &4.243 + 0.07442 \times 0 + 0.2811 \times 0 - 0.3416 \times 1 + 0.4085 \times 0 + 0.03714 \times 0 \\ &+ 0.2226 \times 0 - 0.09804 \times (2023 - 2013) - 5.482 \times 10^{-7} \times 50000 + 0.4685 \times 0 \\ &- 0.08021 \times 0 - 0.1808 \times 0 - 0.2075 \times 0 + 0.06682 \times 0 + 0.5092 \times 0 \\ &+ 0.8702 \times 0 - 0.06067 \times 0 - 0.1631 \times 0 - 0.1322 \times 0 + 0.2984 \times 0 \\ \mu = &18.05802\end{aligned}$$

Jadi, dengan model yang diperoleh pada karya tulis ini, orang tersebut disarankan untuk menjual mobil bekasnya seharga 180580.2 INR.

Daftar Pustaka

Jong, P. D., & Heller, G. Z. (2008). Generalized linear models for insurance data. Cambridge University Press.matheco.2012.04.001

Goldburd, M., Khare, A., & Tevet, D. (2016). Generalized linear models for insurance rating.

Vehicle Dataset from Cardekho. (n.d.). Kaggle. Retrieved April 24, 2023, from https://www.kaggle.com/datasets/nehalbirla/vehicle-dataset-from-cardekho

A Lampiran

Berikut merupakan *code* pemrograman R yang digunakan dalam penyusunan karya tulis ini:

Lampiran Pemrograman R

10820033 Pamella Cathryn

```
library(readxl) # membaca file csv
library(MASS) # regresi
library(glm2) # lebih stabil daripada fungsi glm
library(fitdistrplus) # estimasi parameter
library(ggplot2) # visualisasi
library(dplyr) # visualisasi
library(actuar) # distribusi inverse gaussian
library(Metrics) # menghitung MAPE

# import dataframe
df_full <- read.csv('data_car_new.csv')

# assign variabel kategorikal
df_full$brand = as.factor(df_full$brand)
df_full$fuel = as.factor(df_full$fuel)
df_full$seller_type = as.factor(df_full$seller_type)
df_full$transmission = as.factor(df_full$transmission)
df_full$owner = as.factor(df_full$owner)

# base level variabel kategorikal
df_full = df_full %>% mutate(brand = relevel(brand, ref = "Maruti"))
df_full = df_full %>% mutate(fuel = relevel(fuel, ref = "Petrol"))
df_full = df_full %>% mutate(seller_type = relevel(seller_type, ref =
"Individual"))
df_full = df_full %>% mutate(transmission = relevel(transmission, ref =
"Manual"))
df_full = df_full %>% mutate(owner = relevel(owner, ref = "First Owner"))

# cuplikan data
head(df_full)
```

```
##   selling_price   brand year km_driven   fuel seller_type transmission
## 1           6.0  Maruti   16     70000 Petrol  Individual       Manual
## 2          13.5  Maruti   16     50000 Petrol  Individual       Manual
## 3          60.0 Hyundai   11    100000 Diesel  Individual       Manual
## 4          25.0  others    6     46000 Petrol  Individual       Manual
## 5          45.0   Honda    9    141000 Diesel  Individual       Manual
## 6          14.0  Maruti   16    125000 Petrol  Individual       Manual
##          owner
## 1  First Owner
## 2  First Owner
## 3  First Owner
## 4  First Owner
```

```
## 5 Second Owner
## 6  First Owner
```

```
# randomize data
set.seed(220602)
df_randomized <- df_full[sample(1:nrow(df_full)),]

# memisahkan data train dan data test
sep_index = seq(1, nrow(df_randomized)*0.8,1)
df = df_randomized[sep_index,]
df_test = df_randomized[-sep_index,]

# statistika deskriptif
summary(df)
```

```
##  selling_price         brand          year          km_driven
##  Min.   :  2.00   Maruti  :995   Min.   : 3.000   Min.   :     1
##  1st Qu.: 20.07   Ford    :194   1st Qu.: 7.000   1st Qu.: 35000
##  Median : 35.00   Honda   :208   Median : 9.000   Median : 60000
##  Mean   : 50.08   Hyundai :671   Mean   : 9.906   Mean   : 66828
##  3rd Qu.: 60.00   Mahindra:290   3rd Qu.:12.000   3rd Qu.: 90000
##  Max.   :890.00   others  :818   Max.   :31.000   Max.   :806599
##                   Tata    :296
##       fuel               seller_type      transmission
##  Petrol  :1717   Individual      :2602   Manual   :3108
##  CNG     :  30   Dealer          : 790   Automatic: 364
##  Diesel  :1705   Trustmark Dealer:  80
##  Electric:   1
##  LPG     :  19
##
##
##                    owner
##  First Owner         :2243
##  Fourth & Above Owner:  68
##  Second Owner        : 912
##  Test Drive Car      :  14
##  Third Owner         : 235
##
##
```

```
str(df)
```

```
## 'data.frame':    3472 obs. of  8 variables:
##  $ selling_price: num  12 16 49 48 52.5 ...
##  $ brand        : Factor w/ 7 levels "Maruti","Ford",..: 7 4 4 5 4 7 7 1 4
4 ...
##  $ year         : int  12 10 6 13 9 12 11 15 11 11 ...
##  $ km_driven    : int  20000 80000 66000 90000 54000 20000 80000 70000
49824 40000 ...
##  $ fuel         : Factor w/ 5 levels "Petrol","CNG",..: 1 1 3 3 1 1 3 1 1
1 ...
```

```
##  $ seller_type  : Factor w/ 3 levels "Individual","Dealer",..: 1 1 2 1 2 1
1 1 2 1 ...
##  $ transmission : Factor w/ 2 levels "Manual","Automatic": 1 1 1 1 1 1 1 1
1 1 ...
##  $ owner        : Factor w/ 5 levels "First Owner",..: 1 1 1 1 3 1 1 5 1 1
...
```

```r
y <- df$selling_price
x1 <- df$brand
x2 <- df$year
x3 <- df$km_driven
x4 <- df$fuel
x5 <- df$seller_type
x6 <- df$transmission
x7 <- df$owner

# gambar-gambar EDA
hist(y, breaks = 50, main = "Histogram Selling Price", xlab = "Selling Price
(10000 INR)")

hist(x2, breaks = 50, main = "Histogram Year", xlab = "Year")

hist(x3, breaks = 50, main = "Histogram Km Driven", xlab = "Distance (km)")

data <- x1
freq_table <- table(data)
percentages <- round(prop.table(freq_table) * 100, 1)
par(mar = c(1, 1, 1, 1))
pie(freq_table,
    main = "Pie Chart Brand",
    col = rainbow(length(freq_table)),
    labels = paste(names(freq_table), ":", percentages, "%"))

data <- x4
freq_table <- table(data)
percentages <- round(prop.table(freq_table) * 100, 1)
par(mar = c(1, 1, 1, 1))
pie(freq_table,
    main = "Pie Chart Fuel",
    col = rainbow(length(freq_table)),
    labels = paste(names(freq_table), ":", percentages, "%"))

data <- x5
freq_table <- table(data)
percentages <- round(prop.table(freq_table) * 100, 1)
par(mar = c(1, 1, 1, 1))
pie(freq_table,
    main = "Pie Chart Seller Type",
    col = rainbow(length(freq_table)),
    labels = paste(names(freq_table), ":", percentages, "%"))
```

```r
data <- x6
freq_table <- table(data)
percentages <- round(prop.table(freq_table) * 100, 1)
par(mar = c(1, 1, 1, 1))
pie(freq_table,
    main = "Pie Chart Transmission",
    col = rainbow(length(freq_table)),
    labels = paste(names(freq_table), ":", percentages, "%"))

data <- x7
freq_table <- table(data)
percentages <- round(prop.table(freq_table) * 100, 1)
par(mar = c(1, 1, 1, 1))
pie(freq_table,
    main = "Pie Chart Owner",
    col = rainbow(length(freq_table)),
    labels = paste(names(freq_table), ":", percentages, "%"))

order <- c("Maruti", "Hyundai", "Mahindra", "Tata", "Honda", "Ford",
"others")
x1 <- factor(x1, levels = order)
ggplot(df, aes(x = x1, y = y, fill = x1)) +
  geom_boxplot() +
  labs(x = "Brand", y = "Selling Price (10000 INR)", title = "Box Plot Harga
Mobil Bekas Berdasarkan Merek") +
  theme_light() +
  guides(fill = FALSE)

df %>%
  ggplot() +
  aes(x = x2, y = y) +
  geom_point(size = 1L, col="#877BAE") +
  labs(x = "Year", y = "Selling Price (10000 INR)", title = "Scatter Plot
Harga Mobil Bekas Terhadap Tahun") +
  theme_light() +
  guides(fill = FALSE)

df %>%
  ggplot() +
  aes(x = x3, y = y) +
  geom_point(size = 1L, col="#877BAE") +
  labs(x = "Distance (km)", y = "Selling Price (10000 INR)", title = "Scatter
Plot Harga Mobil Bekas Terhadap Jarak Tempuh Mobil") +
  theme_light() +
  guides(fill = FALSE)

order <- c("Petrol", "Diesel", "CNG", "LPG", "Electric")
x4 <- factor(x4, levels = order)
ggplot(df, aes(x = x4, y = y, fill = x4)) +
  geom_boxplot() +
  labs(x = "Fuel", y = "Selling Price (10000 INR)", title = "Box Plot Harga
```

```
Mobil Bekas Berdasarkan Jenis Bahan Bakar") +
  theme_light() +
  guides(fill = FALSE)

order <- c("Individual", "Dealer", "Trustmark Dealer")
x5 <- factor(x5, levels = order)
ggplot(df, aes(x = x5, y = y, fill = x5)) +
  geom_boxplot() +
  labs(x = "Seller Type", y = "Selling Price (10000 INR)", title = "Box Plot
Harga Mobil Bekas Berdasarkan Tipe Penjual") +
  theme_light() +
  guides(fill = FALSE)

order <- c("Manual", "Automatic")
x6 <- factor(x6, levels = order)
ggplot(df, aes(x = x6, y = y, fill = x6)) +
  geom_boxplot() +
  labs(x = "Transmission", y = "Selling Price (10000 INR)", title = "Box Plot
Harga Mobil Bekas Berdasarkan Transmisinya") +
  theme_light() +
  guides(fill = FALSE)

order <- c("First Owner", "Second Owner", "Third Owner", "Fourth & Above
Owner", "Test Drive Car")
x7 <- factor(x7, levels = order)
ggplot(df, aes(x = x7, y = y, fill = x7)) +
  geom_boxplot() +
  labs(x = "Owner", y = "Selling Price (10000 INR)", title = "Box Plot Harga
Mobil Bekas Berdasarkan Jumlah Pemilik Sebelumnya") +
  theme_light() +
  guides(fill = FALSE)

# gambar-gambar hubungan
df %>%
 filter(brand %in% "Maruti") %>%
 ggplot() +
 aes(x = km_driven, y = selling_price, colour = brand) +
 geom_point(shape = "circle", size = 2L) +  geom_smooth(method = "glm", se =
FALSE) +
 scale_color_manual(values = c(Maruti = "#A6CEE3", Ford = "#99CD91", Honda =
"#B89B74", Hyundai = "#F06C45",
 Mahindra = "#ED8F47", others = "#825D99", Tata = "#B15928")) +
 labs(x = "Distance (km)", y = "Selling Price (10000 INR)",
 title = "Scatter Plot Harga Mobil Bekas Terhadap Jarak Tempuh Mobil untuk
Merek Maruti") +
 theme_minimal() +
 theme(legend.position = "none")

## `geom_smooth()` using formula = 'y ~ x'
```

```
df %>%
 filter(brand %in% "Hyundai") %>%
 ggplot() +
 aes(x = km_driven, y = selling_price, colour = brand) +
 geom_point(shape = "circle", size = 2L) +  geom_smooth(method = "glm", se =
FALSE) +
 scale_color_manual(values = c(Maruti = "#A6CEE3", Ford = "#99CD91", Honda =
"#B89B74", Hyundai = "#F06C45",
 Mahindra = "#ED8F47", others = "#825D99", Tata = "#B15928")) +
 labs(x = "Distance (km)", y = "Selling Price (10000 INR)",
 title = "Scatter Plot Harga Mobil Bekas Terhadap Jarak Tempuh Mobil untuk
Merek Hyundai") +
 theme_minimal() +
 theme(legend.position = "none")

## `geom_smooth()` using formula = 'y ~ x'
```

```
df %>%
 filter(brand %in% "Mahindra") %>%
 ggplot() +
 aes(x = km_driven, y = selling_price, colour = brand) +
 geom_point(shape = "circle", size = 2L) +  geom_smooth(method = "glm", se =
FALSE) +
 scale_color_manual(values = c(Maruti = "#A6CEE3", Ford = "#99CD91", Honda =
"#B89B74", Hyundai = "#F06C45",
 Mahindra = "#ED8F47", others = "#825D99", Tata = "#B15928")) +
 labs(x = "Distance (km)", y = "Selling Price (10000 INR)",
 title = "Scatter Plot Harga Mobil Bekas Terhadap Jarak Tempuh Mobil untuk
Merek Mahindra") +
 theme_minimal() +
 theme(legend.position = "none")

## `geom_smooth()` using formula = 'y ~ x'
```

```
df %>%
 filter(brand %in% "Tata") %>%
 ggplot() +
 aes(x = km_driven, y = selling_price, colour = brand) +
 geom_point(shape = "circle", size = 2L) +  geom_smooth(method = "glm", se =
FALSE) +
 scale_color_manual(values = c(Maruti = "#A6CEE3", Ford = "#99CD91", Honda =
"#B89B74", Hyundai = "#F06C45",
 Mahindra = "#ED8F47", others = "#825D99", Tata = "#B15928")) +
 labs(x = "Distance (km)", y = "Selling Price (10000 INR)",
 title = "Scatter Plot Harga Mobil Bekas Terhadap Jarak Tempuh Mobil untuk
Merek Tata") +
 theme_minimal() +
 theme(legend.position = "none")

## `geom_smooth()` using formula = 'y ~ x'
```

```
df %>%
 filter(brand %in% "Honda") %>%
 ggplot() +
 aes(x = km_driven, y = selling_price, colour = brand) +
 geom_point(shape = "circle", size = 2L) +  geom_smooth(method = "glm", se =
FALSE) +
 scale_color_manual(values = c(Maruti = "#A6CEE3", Ford = "#99CD91", Honda =
"#B89B74", Hyundai = "#F06C45",
 Mahindra = "#ED8F47", others = "#825D99", Tata = "#B15928")) +
 labs(x = "Distance (km)", y = "Selling Price (10000 INR)",
 title = "Scatter Plot Harga Mobil Bekas Terhadap Jarak Tempuh Mobil untuk
Merek Honda") +
 theme_minimal() +
 theme(legend.position = "none")

## `geom_smooth()` using formula = 'y ~ x'
```

```
df %>%
 filter(brand %in% "Ford") %>%
 ggplot() +
 aes(x = km_driven, y = selling_price, colour = brand) +
 geom_point(shape = "circle", size = 2L) +  geom_smooth(method = "glm", se =
FALSE) +
 scale_color_manual(values = c(Maruti = "#A6CEE3", Ford = "#99CD91", Honda =
"#B89B74", Hyundai = "#F06C45",
 Mahindra = "#ED8F47", others = "#825D99", Tata = "#B15928")) +
 labs(x = "Distance (km)", y = "Selling Price (10000 INR)",
 title = "Scatter Plot Harga Mobil Bekas Terhadap Jarak Tempuh Mobil untuk
Merek Ford") +
 theme_minimal() +
 theme(legend.position = "none")

## `geom_smooth()` using formula = 'y ~ x'
```

```
df %>%
 filter(brand %in% "others") %>%
 ggplot() +
 aes(x = km_driven, y = selling_price, colour = brand) +
 geom_point(shape = "circle", size = 2L) +  geom_smooth(method = "glm", se =
FALSE) +
 scale_color_manual(values = c(Maruti = "#A6CEE3", Ford = "#99CD91", Honda =
"#B89B74", Hyundai = "#F06C45",
 Mahindra = "#ED8F47", others = "#825D99", Tata = "#B15928")) +
 labs(x = "Distance (km)", y = "Selling Price (10000 INR)",
 title = "Scatter Plot Harga Mobil Bekas Terhadap Jarak Tempuh Mobil untuk
Merek Lainnya") +
 theme_minimal() +
 theme(legend.position = "none")

## `geom_smooth()` using formula = 'y ~ x'
```

```r
# memeriksa multikolinearitas
corr <- data.frame(x2, x3)
(round(cor(corr),2))
```

```
##      x2   x3
## x2 1.00 0.42
## x3 0.42 1.00
```

```r
# distribusi gamma
set.seed(220602)

data <- y

fit_gamma <- fitdist(data, "gamma", method = "mle")
fit_gamma$estimate
```

```
##      shape       rate
## 1.82261229 0.04131099
```

```r
shape_gamma <- fit_gamma$estimate[1]
rate_gamma <- fit_gamma$estimate[2]
ks.test(data, rgamma(nrow(df), shape = shape_gamma, rate = rate_gamma))
```

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  data and rgamma(nrow(df), shape = shape_gamma, rate = rate_gamma)
## D = 0.014562, p-value = 0.1201
## alternative hypothesis: two-sided
```

```r
cat("AIC:", summary(fit_gamma)$aic, "\n")
```

```
## AIC: 33971.99
```

```r
cat("BIC:", summary(fit_gamma)$bic, "\n")
```

```
## BIC: 33984.29
```

```r
h = hist(data, freq = FALSE, breaks = 50, main="Histogram Selling Price
dengan Kurva Gamma", xlab="Selling Price (10000 INR)")
curve(dgamma(x, shape = shape_gamma, rate = rate_gamma), col="red", add=TRUE)

# distribusi inverse gaussian
set.seed(220602)

data <- y

fit_invgauss <- fitdist(data, "invgauss", method = "mle", lower=c(0,0), start
= list(mean = mean(data), shape = mean(data)^3/sd(data)^2))
fit_invgauss$estimate
```

```
##     mean    shape
## 48.63472 58.86465

mean_invgauss <- fit_invgauss$estimate[1]
shape_invgauss <- fit_invgauss$estimate[2]
ks.test(data, rinvgauss(nrow(df), mean = mean_invgauss, shape =
shape_invgauss))

##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  data and rinvgauss(nrow(df), mean = mean_invgauss, shape =
shape_invgauss)
## D = 0.027442, p-value = 0.0803
## alternative hypothesis: two-sided

cat("AIC:", summary(fit_invgauss)$aic, "\n")

## AIC: 33390.71

cat("BIC:", summary(fit_invgauss)$bic, "\n")

## BIC: 33403.02

h = hist(data, freq = FALSE, breaks = 50, main="Histogram Selling Price
dengan Kurva Inverse Gaussian", xlab="Selling Price (10000 INR)")
curve(dinvgauss(x, mean = mean_invgauss, shape = shape_invgauss), col="red",
add=TRUE)

# Model 1
model1 <- glm2(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, family = Gamma(link =
"inverse"), start = rep(1,20), data = df)
summary(model1)

##
## Call:
## glm2(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, family = Gamma(link =
"inverse"),
##     data = df, start = rep(1, 20))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4122  -0.4244  -0.1218   0.1864   4.5156
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.426e-02  7.973e-04  17.883  < 2e-16 ***
## x1Hyundai              -3.624e-03  7.288e-04  -4.972 6.95e-07 ***
## x1Mahindra             -5.124e-03  7.848e-04  -6.530 7.55e-11 ***
## x1Tata                  5.357e-03  1.312e-03   4.084 4.52e-05 ***
## x1Honda                -6.005e-03  8.869e-04  -6.771 1.50e-11 ***
## x1Ford                 -4.239e-03  8.576e-04  -4.943 8.04e-07 ***
```

```
## x1others                -5.209e-03  6.580e-04  -7.916 3.27e-15 ***
## x2                       2.120e-03  7.451e-05  28.461  < 2e-16 ***
## x3                      -6.178e-09  3.150e-09  -1.961 0.049933 *
## x4Diesel                -8.169e-03  5.019e-04 -16.275  < 2e-16 ***
## x4CNG                    5.052e-03  4.084e-03   1.237 0.216163
## x4LPG                    1.797e-02  7.946e-03   2.262 0.023758 *
## x4Electric              -5.341e-03  1.958e-02  -0.273 0.785033
## x5Dealer                -1.391e-03  2.507e-04  -5.549 3.08e-08 ***
## x5Trustmark Dealer      -3.704e-03  5.383e-04  -6.881 7.02e-12 ***
## x6Automatic             -6.891e-03  4.586e-04 -15.026  < 2e-16 ***
## x7Second Owner           1.041e-03  6.132e-04   1.697 0.089733 .
## x7Third Owner            5.430e-03  1.474e-03   3.683 0.000234 ***
## x7Fourth & Above Owner   1.679e-02  4.000e-03   4.198 2.76e-05 ***
## x7Test Drive Car         3.490e-04  1.747e-03   0.200 0.841643
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.4179357)
##
##     Null deviance: 2521.56  on 3471  degrees of freedom
## Residual deviance:  908.71  on 3452  degrees of freedom
## AIC: 30047
##
## Number of Fisher Scoring iterations: 25

y_pred_1 <- 1/(predict(model1, df))
cat("MAPE:", mape(df$selling_price, y_pred_1), "\n")

## MAPE: 0.514983

# Model 2
model2 <- glm2(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, family =
inverse.gaussian(link = "log"), data = df)
summary(model2)

##
## Call:
## glm2(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, family =
inverse.gaussian(link = "log"),
##     data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.58399  -0.06428  -0.01285   0.02975   0.80529
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              4.243e+00  3.157e-02 134.388  < 2e-16 ***
## x1Hyundai                7.442e-02  2.452e-02   3.035  0.00242 **
## x1Mahindra               2.811e-01  4.452e-02   6.315 3.04e-10 ***
## x1Tata                  -3.416e-01  3.029e-02 -11.279  < 2e-16 ***
```

```
## x1Honda                  4.085e-01  4.715e-02   8.665  < 2e-16 ***
## x1Ford                   3.714e-02  4.502e-02   0.825  0.40943
## x1others                 2.226e-01  2.872e-02   7.750 1.20e-14 ***
## x2                      -9.804e-02  2.126e-03 -46.126  < 2e-16 ***
## x3                      -5.482e-07  2.031e-07  -2.699  0.00699 **
## x4Diesel                 4.685e-01  2.292e-02  20.442  < 2e-16 ***
## x4CNG                   -8.021e-02  8.128e-02  -0.987  0.32378
## x4LPG                   -1.808e-01  8.289e-02  -2.181  0.02924 *
## x4Electric              -2.075e-01  5.320e-01  -0.390  0.69656
## x5Dealer                 6.682e-02  2.593e-02   2.577  0.01001 *
## x5Trustmark Dealer       5.092e-01  9.670e-02   5.265 1.49e-07 ***
## x6Automatic              8.702e-01  5.264e-02  16.530  < 2e-16 ***
## x7Second Owner          -6.067e-02  2.137e-02  -2.839  0.00455 **
## x7Third Owner           -1.631e-01  3.145e-02  -5.187 2.26e-07 ***
## x7Fourth & Above Owner -1.322e-01  4.817e-02  -2.744  0.00611 **
## x7Test Drive Car         2.984e-01  2.522e-01   1.183  0.23670
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for inverse.gaussian family taken to be 0.009014588)
##
##     Null deviance: 70.829  on 3471  degrees of freedom
## Residual deviance: 24.246  on 3452  degrees of freedom
## AIC: 29641
##
## Number of Fisher Scoring iterations: 18

y_pred_2 <- exp(predict(model2, df))
cat("MAPE:", mape(df$selling_price, y_pred_2), "\n")

## MAPE: 0.3893762

# Model 3
model3 <- glm2(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, family = Gamma(link =
"log"), data = df)
summary(model3)

##
## Call:
## glm2(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, family = Gamma(link =
"log"),
##     data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.33900  -0.37457  -0.08351   0.19824   3.07654
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              4.305e+00  2.934e-02 146.717  < 2e-16 ***
## x1Hyundai                1.172e-01  2.506e-02   4.677 3.03e-06 ***
```

```
## x1Mahindra               2.342e-01  3.562e-02   6.576 5.57e-11 ***
## x1Tata                  -2.890e-01  3.381e-02  -8.546  < 2e-16 ***
## x1Honda                  3.095e-01  3.840e-02   8.059 1.05e-15 ***
## x1Ford                   9.091e-02  4.107e-02   2.214  0.02692 *
## x1others                 2.734e-01  2.529e-02  10.810  < 2e-16 ***
## x2                      -1.065e-01  2.517e-03 -42.301  < 2e-16 ***
## x3                      -4.592e-07  2.145e-07  -2.141  0.03238 *
## x4Diesel                 4.846e-01  2.020e-02  23.990  < 2e-16 ***
## x4CNG                   -6.362e-02  9.205e-02  -0.691  0.48953
## x4LPG                   -1.447e-01  1.151e-01  -1.257  0.20897
## x4Electric              -1.604e-01  4.991e-01  -0.321  0.74796
## x5Dealer                 8.695e-02  2.192e-02   3.966 7.46e-05 ***
## x5Trustmark Dealer       4.108e-01  5.756e-02   7.137 1.16e-12 ***
## x6Automatic              8.255e-01  2.939e-02  28.085  < 2e-16 ***
## x7Second Owner          -5.742e-02  2.155e-02  -2.664  0.00776 **
## x7Third Owner           -1.478e-01  3.666e-02  -4.033 5.63e-05 ***
## x7Fourth & Above Owner -1.737e-01  6.333e-02  -2.742  0.00613 **
## x7Test Drive Car         1.519e-01  1.374e-01   1.105  0.26904
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.2474761)
##
##     Null deviance: 2521.56  on 3471  degrees of freedom
## Residual deviance:  685.17  on 3452  degrees of freedom
## AIC: 29029
##
## Number of Fisher Scoring iterations: 7

y_pred_3 <- exp(predict(model3, df))
cat("MAPE:", mape(df$selling_price, y_pred_3), "\n")

## MAPE: 0.3737762

# Model 4
model4 <- glm2(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x2*x3 + x2*x6 + x2*x7 +
x3*x7, family = Gamma(link = "inverse"), data = df)
summary(model4)

##
## Call:
## glm2(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x2 * x3 +
##     x2 * x6 + x2 * x7 + x3 * x7, family = Gamma(link = "inverse"),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7319  -0.3870  -0.0920   0.2022   3.7901
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)                8.250e-03  8.875e-04   9.296  < 2e-16 ***
## x1Hyundai                 -2.955e-03  5.928e-04  -4.985 6.50e-07 ***
## x1Mahindra                -4.293e-03  6.312e-04  -6.801 1.22e-11 ***
## x1Tata                     4.175e-03  1.078e-03   3.872 0.000110 ***
## x1Honda                   -4.986e-03  7.387e-04  -6.749 1.73e-11 ***
## x1Ford                    -3.635e-03  7.125e-04  -5.102 3.55e-07 ***
## x1others                  -5.839e-03  5.400e-04 -10.814  < 2e-16 ***
## x2                         2.810e-03  1.171e-04  23.992  < 2e-16 ***
## x3                        -1.017e-08  1.032e-08  -0.986 0.324084
## x4Diesel                  -8.044e-03  4.096e-04 -19.638  < 2e-16 ***
## x4CNG                      4.541e-03  3.344e-03   1.358 0.174510
## x4LPG                      1.264e-02  6.683e-03   1.892 0.058574 .
## x4Electric                 3.927e-03  1.671e-02   0.235 0.814202
## x5Dealer                  -1.310e-03  2.954e-04  -4.436 9.47e-06 ***
## x5Trustmark Dealer        -2.032e-03  4.773e-04  -4.257 2.13e-05 ***
## x6Automatic                6.761e-03  7.617e-04   8.876  < 2e-16 ***
## x7Second Owner            -2.456e-03  1.304e-03  -1.883 0.059805 .
## x7Third Owner             -4.690e-03  3.175e-03  -1.477 0.139724
## x7Fourth & Above Owner     1.190e-02  1.217e-02   0.978 0.328078
## x7Test Drive Car           1.433e-02  1.865e-02   0.768 0.442372
## x2:x3                      2.152e-09  1.310e-09   1.643 0.100443
## x2:x6Automatic            -2.251e-03  1.023e-04 -22.001  < 2e-16 ***
## x2:x7Second Owner          9.505e-04  1.621e-04   5.863 4.98e-09 ***
## x2:x7Third Owner           1.605e-03  3.387e-04   4.738 2.24e-06 ***
## x2:x7Fourth & Above Owner  1.875e-03  8.556e-04   2.192 0.028452 *
## x2:x7Test Drive Car       -3.598e-03  6.195e-03  -0.581 0.561448
## x3:x7Second Owner         -7.060e-08  1.080e-08  -6.536 7.27e-11 ***
## x3:x7Third Owner          -7.765e-08  2.174e-08  -3.572 0.000359 ***
## x3:x7Fourth & Above Owner -2.206e-07  7.034e-08  -3.137 0.001721 **
## x3:x7Test Drive Car        3.740e-07  7.852e-07   0.476 0.633856
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.2659053)
##
##     Null deviance: 2521.56  on 3471  degrees of freedom
## Residual deviance:  741.64  on 3442  degrees of freedom
## AIC: 29334
##
## Number of Fisher Scoring iterations: 6

y_pred_4 <- 1/(predict(model4, df))
cat("MAPE:", mape(df$selling_price, y_pred_4), "\n")

## MAPE: 0.4458053

# Model 5
model5 <- glm2(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x2*x3 + x2*x6 + x2*x7 +
x3*x7, family = inverse.gaussian(link = "log"), data = df)
summary(model5)
```

```
##
## Call:
## glm2(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x2 * x3 +
##     x2 * x6 + x2 * x7 + x3 * x7, family = inverse.gaussian(link = "log"),
##     data = df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.57686  -0.06277  -0.01324   0.02927   0.73349
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  4.391e+00  4.705e-02  93.322  < 2e-16 ***
## x1Hyundai                    8.185e-02  2.372e-02   3.450 0.000566 ***
## x1Mahindra                   2.863e-01  4.332e-02   6.610 4.42e-11 ***
## x1Tata                      -3.391e-01  2.928e-02 -11.580  < 2e-16 ***
## x1Honda                      3.968e-01  4.589e-02   8.648  < 2e-16 ***
## x1Ford                       3.121e-02  4.344e-02   0.718 0.472560
## x1others                     2.265e-01  2.780e-02   8.147 5.16e-16 ***
## x2                          -1.130e-01  3.938e-03 -28.688  < 2e-16 ***
## x3                          -6.727e-07  6.365e-07  -1.057 0.290622
## x4Diesel                     4.736e-01  2.245e-02  21.098  < 2e-16 ***
## x4CNG                       -6.813e-02  7.863e-02  -0.866 0.386312
## x4LPG                       -1.739e-01  8.070e-02  -2.155 0.031206 *
## x4Electric                  -3.634e-01  5.200e-01  -0.699 0.484723
## x5Dealer                     7.102e-02  2.508e-02   2.832 0.004654 **
## x5Trustmark Dealer           4.738e-01  9.335e-02   5.076 4.06e-07 ***
## x6Automatic                  5.523e-01  1.101e-01   5.014 5.59e-07 ***
## x7Second Owner              -3.543e-01  6.649e-02  -5.329 1.05e-07 ***
## x7Third Owner               -3.878e-01  1.229e-01  -3.155 0.001618 **
## x7Fourth & Above Owner      -9.944e-01  1.817e-01  -5.473 4.75e-08 ***
## x7Test Drive Car            -8.603e-01  3.159e+00  -0.272 0.785405
## x2:x3                        1.887e-08  4.810e-08   0.392 0.694876
## x2:x6Automatic               2.833e-02  1.028e-02   2.757 0.005864 **
## x2:x7Second Owner            1.874e-02  4.446e-03   4.216 2.55e-05 ***
## x2:x7Third Owner             2.280e-02  6.893e-03   3.308 0.000948 ***
## x2:x7Fourth & Above Owner    4.648e-02  8.286e-03   5.609 2.19e-08 ***
## x2:x7Test Drive Car          3.775e-01  1.068e+00   0.353 0.723822
## x3:x7Second Owner            8.535e-07  4.622e-07   1.847 0.064905 .
## x3:x7Third Owner            -6.726e-07  5.853e-07  -1.149 0.250606
## x3:x7Fourth & Above Owner    1.136e-06  1.245e-06   0.913 0.361219
## x3:x7Test Drive Car         -6.383e-05  1.308e-04  -0.488 0.625523
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for inverse.gaussian family taken to be 0.008429643)
##
##     Null deviance: 1873.92  on 3471  degrees of freedom
## Residual deviance: 768.03  on 3442  degrees of freedom
## AIC: 29566
```

```
##
## Number of Fisher Scoring iterations: 25

y_pred_5 <- exp(predict(model5, df))
cat("MAPE:", mape(df$selling_price, y_pred_5), "\n")

## MAPE: 0.9554728

# Model 6
model6 <- glm2(y ~ x1 + offset(log(x2)) + offset(log(x3)) + x4 + x5 + x6 +
x7, family = Gamma(link = "log"), data = df)
summary(model6)

##
## Call:
## glm2(formula = y ~ x1 + offset(log(x2)) + offset(log(x3)) + x4 +
##     x5 + x6 + x7, family = Gamma(link = "log"), data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.9606  -1.8180  -1.2856  -0.5211  23.6047
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -8.03576    0.23319 -34.459  < 2e-16 ***
## x1Hyundai                0.26904    0.30402   0.885  0.37625
## x1Mahindra               4.62191    0.43245  10.688  < 2e-16 ***
## x1Tata                  -0.15608    0.41049  -0.380  0.70381
## x1Honda                  0.72232    0.46597   1.550  0.12120
## x1Ford                   0.81235    0.49847   1.630  0.10326
## x1others                 0.55811    0.30691   1.818  0.06908 .
## x4Diesel                -0.06257    0.23008  -0.272  0.78568
## x4CNG                   -0.34479    1.11621  -0.309  0.75742
## x4LPG                   -2.02829    1.39686  -1.452  0.14658
## x4Electric              -3.00219    6.05776  -0.496  0.62021
## x5Dealer                 0.13569    0.26539   0.511  0.60917
## x5Trustmark Dealer       0.20073    0.69670   0.288  0.77327
## x6Automatic              1.06531    0.35329   3.015  0.00259 **
## x7Second Owner          -1.15528    0.24369  -4.741 2.21e-06 ***
## x7Third Owner           -2.85170    0.42093  -6.775 1.46e-11 ***
## x7Fourth & Above Owner  -0.96848    0.74885  -1.293  0.19600
## x7Test Drive Car         4.52960    1.66128   2.727  0.00643 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 36.4914)
##
##     Null deviance: 20835  on 3471  degrees of freedom
## Residual deviance: 11703  on 3454  degrees of freedom
## AIC: 40418
```

```
##
## Number of Fisher Scoring iterations: 25

y_pred_6 <- exp(predict(model6, df))
cat("MAPE:", mape(df$selling_price, y_pred_6), "\n")

## MAPE: 51.00181

# Model 7
model7 <- glm2(y ~ x1 + offset(log(x2)) + offset(log(x3)) + x4 + x5 + x6 +
x7, family = inverse.gaussian(link = "log"), data = df)
summary(model7)

##
## Call:
## glm2(formula = y ~ x1 + offset(log(x2)) + offset(log(x3)) + x4 +
##     x5 + x6 + x7, family = inverse.gaussian(link = "log"), data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6673  -0.2066  -0.1528  -0.1123   0.5996
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -4.995865   0.066302 -75.350  < 2e-16 ***
## x1Hyundai                1.751831   0.077158  22.705  < 2e-16 ***
## x1Mahindra              24.245614 600.565771   0.040   0.9678
## x1Tata                  -0.648124   0.043427 -14.924  < 2e-16 ***
## x1Honda                 -0.112915   0.077401  -1.459   0.1447
## x1Ford                   0.007716   0.071142   0.108   0.9136
## x1others                 1.779201   0.077712  22.895  < 2e-16 ***
## x4Diesel                 0.123640   0.038415   3.219   0.0013 **
## x4CNG                   -1.153174   0.114657 -10.058  < 2e-16 ***
## x4LPG                   -1.721563   0.090750 -18.970  < 2e-16 ***
## x4Electric              -6.768963   0.445349 -15.199  < 2e-16 ***
## x5Dealer                 1.763919   0.100849  17.491  < 2e-16 ***
## x5Trustmark Dealer      -1.912416   0.123625 -15.469  < 2e-16 ***
## x6Automatic              2.282418   0.179634  12.706  < 2e-16 ***
## x7Second Owner          -4.494826   0.066566 -67.524  < 2e-16 ***
## x7Third Owner           -5.196327   0.073066 -71.118  < 2e-16 ***
## x7Fourth & Above Owner  -0.039449   0.313519  -0.126   0.8999
## x7Test Drive Car         1.632522   0.331043   4.931 8.55e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for inverse.gaussian family taken to be 0.005103683)
##
##     Null deviance: 140.16  on 3471  degrees of freedom
## Residual deviance: 122.55  on 3454  degrees of freedom
## AIC: 35263
```

```
##
## Number of Fisher Scoring iterations: 25

y_pred_7 <- exp(predict(model7, df))
cat("MAPE:", mape(df$selling_price, y_pred_7), "\n")

## MAPE: 45.63847

# deviance residuals
dev_resid_2 <- residuals(model2, type = "deviance")
summary(dev_resid_2)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.58399 -0.06428 -0.01285 -0.01660  0.02975  0.80529

dev_resid_3 <- residuals(model3, type = "deviance")
summary(dev_resid_3)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1.33900 -0.37457 -0.08351 -0.06574  0.19824  3.07654

residual <- data.frame(1:nrow(df), dev_resid_2, dev_resid_3)

ggplot(residual) +
 aes(x = X1.nrow.df., y = dev_resid_2) +
 geom_point(shape = "circle", size = 1.5, colour = "#112446") +
 labs(x = "Observasi ke-", y = "Deviance Residuals", title = "Deviance
Residuals Model 2") +
 theme_bw()

ggplot(residual) +
 aes(x = X1.nrow.df., y = dev_resid_3) +
 geom_point(shape = "circle", size = 1.5, colour = "#112446") +
 labs(x = "Observasi ke-", y = "Deviance Residuals", title = "Deviance
Residuals Model 3") +
 theme_bw()

# Uji Data Test Model 2
y_actual_test = df_test$selling_price
y_pred_test = exp(predict(model2, df_test))
cat("MAPE:", mape(y_actual_test, y_pred_test), "\n")

## MAPE: 0.416318

# Uji Data Test Model 3
y_actual_test = df_test$selling_price
y_pred_test = exp(predict(model3, df_test))
cat("MAPE:", mape(y_actual_test, y_pred_test), "\n")

## MAPE: 0.551836
```