Analisis Regresi terkait Biaya Premi Asuransi Menggunakan Model Regresi Linier Berganda

Kelompok A

1. Angelica Daphne 2206031441
2. Azriel Graciosa 2206051916
3. Braneal Obed 2206051866
4. Hasnabilla Chandra 2206811902
5. Reinaldus Kevin 2206051683

LATAR BELAKANG

- Premi ditentukan dari berbagai faktor yang diperoleh melalui underwriting.
- Faktor yang mempengaruhi:
  1. Usia,
  2. Jenis kelamin,
  3. BMI,
  4. Jumlah anak,
  5. Tempat tinggal,
  6. Apakah individu memiliki kebiasaan seperti merokok.
- Perlu adanya pemodelan untuk memperkirakan premi yang sesuai untuk seorang individu dimana pemodelan tersebut diharapkan dapat menentukan kisaran premi yang sesuai.

RUMUSAN MASALAH DAN TUJUAN

RUMUSAN MASALAH

1. Bagaimana model yang terbaik untuk menentukan pengaruh variabel-variabel prediktor (usia, indeks massa tubuh, jenis kelamin, tempat tinggal, jumlah anak, dan konsumsi rokok) terhadap besarnya tagihan asuransi?
2. Variabel prediktor apa saja yang memiliki pengaruh signifikan terhadap besarnya tagihan asuransi?

TUJUAN

1. Mengetahui model yang terbaik untuk menentukan variabel-variabel prediktor yang paling berpengaruh pada besarnya tagihan asuransi
2. Mengetahui variabel-variabel yang memiliki pengaruh paling signifikan terhadap besarnya tagihan asuransi.

DESKRIPSI DATA DAN VARIABEL RESPON

Ukuran	: 1338
Jumlah Pengukuran	: 7
Variabel respon	: charges
Variabel prediktor	: usia (*age*), BMI, jenis kelamin (*sex*), perokok atau bukan (*smoker*), banyak anak (*children*), tempat tinggal (*region*)

DESKRIPSI DATA DAN VARIABEL RESPON

Tipe dan Skala Data

1. Age : data numerik skala rasio
2. Sex : data kategorik skala nominal
3. BMI : data numerik skala rasio
4. Children : data numerik skala rasio
5. Smoker : data kategorik skala nominal
6. Region : data kategorik skala nominal
7. Charges : data numerik skala rasio

PRE-PROCESSING

1. Pengecekan data kosong
2. Eliminasi outlier
3. Pengubahan data kategorik menjadi numerik

	age	sex	bmi	children	smoker	region	charges
0	18	male	23.210	0	no	southeast	1121.8739
1	18	male	30.140	0	no	southeast	1131.5066
2	18	male	33.330	0	no	southeast	1135.9407
3	18	male	33.660	0	no	southeast	1136.3994
4	18	male	34.100	0	no	southeast	1137.0110
...	...	...	...	...	...	...	...
1333	51	female	39.500	1	no	southwest	9880.0680
1334	50	female	30.115	1	no	northwest	9910.3599
1335	51	male	27.740	1	no	northeast	9957.7216
1336	51	male	32.300	1	no	northeast	9964.0600
1337	52	female	18.340	0	no	northwest	9991.0377

1338 rows × 7 columns

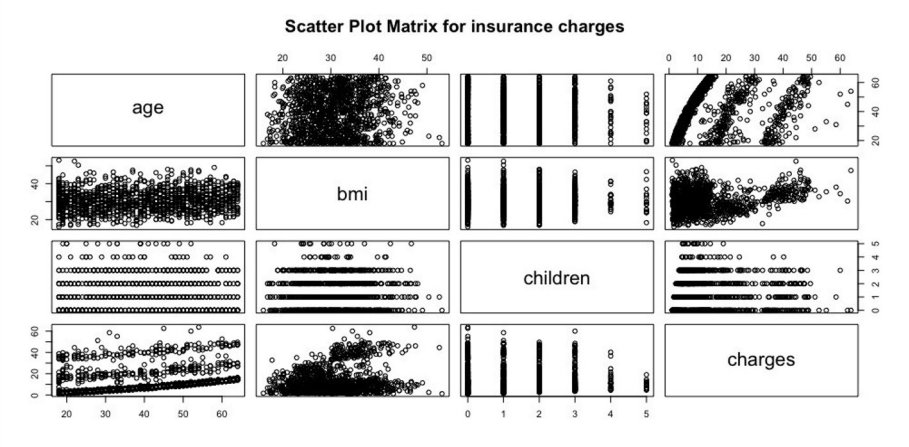
Data sebelum Pre-processing

	age	sex	bmi	children	smoker	charges	northwest	southeast	southwest
0	18	0	23.210	0	0	1121.8739	0	1	0
1	18	0	30.140	0	0	1131.5066	0	1	0
2	18	0	33.330	0	0	1135.9407	0	1	0
3	18	0	33.660	0	0	1136.3994	0	1	0
4	18	0	34.100	0	0	1137.0110	0	1	0
...	...	...	...	...	...	...	...	...	...
1186	51	1	39.500	1	0	9880.0680	0	0	1
1187	50	1	30.115	1	0	9910.3599	1	0	0
1188	51	0	27.740	1	0	9957.7216	0	0	0
1189	51	0	32.300	1	0	9964.0600	0	0	0
1190	52	1	18.340	0	0	9991.0377	1	0	0

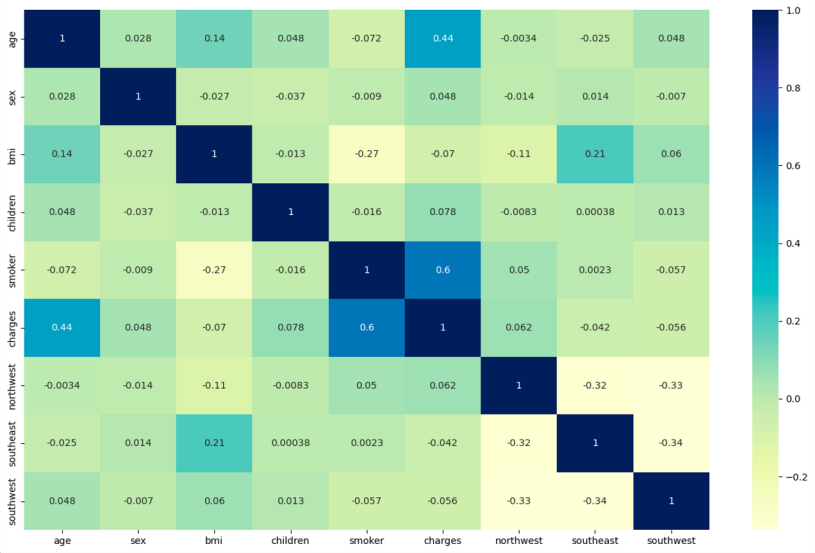
1191 rows × 9 columns

Data setelah Pre-processing

ANALISIS DESKRIPTIF DAN VISUALISASI



Scatter Plot



Heatmap

METODE SELEKSI MODEL

1. Seluruh variabel prediktor nya *statistically useful*, di mana $Pr(> |t|) \leq \alpha = 0.05$,
2. Modelnya useful,
3. Nilai Adjusted R-squared yang cukup tinggi,
4. Model tidak overfitting,
5. Nilai VIF yang rendah untuk setiap variabel prediktor.

PENENTUAN VARIABEL

Variabel prediktor ditentukan menggunakan Recursive Feature Elimination.

1. Age
2. BMI
3. Smoker
4. Children

HIPOTESIS

Biaya premi dipengaruhi oleh minimal satu dari variabel prediktor

ANALISIS MODEL PERTAMA

OLS Regression Results

Dep. Variable:	charges	R-squared:	0.601
Model:	OLS	Adj. R-squared:	0.599
Method:	Least Squares	F-statistic:	311.5
Date:	Fri, 03 Nov 2023	Prob (F-statistic):	2.04e-163
Time:	15:06:00	Log-Likelihood:	458.07
No. Observations:	833	AIC:	-906.1
Df Residuals:	828	BIC:	-882.5
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0251	0.015	1.688	0.092	-0.004	0.054
age	0.3480	0.016	21.719	0.000	0.317	0.379
bmi	0.0412	0.026	1.584	0.114	-0.010	0.092
children	0.0612	0.020	3.005	0.003	0.021	0.101
smoker	0.4368	0.016	28.155	0.000	0.406	0.467

Omnibus:	539.152	Durbin-Watson:	2.112
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3902.374
Skew:	3.063	Prob(JB):	0.00
Kurtosis:	11.655	Cond. No.	7.32

const	9.39
bmi	1.10
smoker	1.08
age	1.02
children	1.00)

$$y = 0.0251 + 0.3480x_1 + 0.0412x_3 + 0.0612x_4 + 0.4368x_5$$

ANALISIS MODEL KEDUA

OLS Regression Results

Dep. Variable:	charges	R-squared:	0.219
Model:	OLS	Adj. R-squared:	0.216
Method:	Least Squares	F-statistic:	77.27
Date:	Tue, 19 Dec 2023	Prob (F-statistic):	4.44e-44
Time:	11:55:32	Log-Likelihood:	178.34
No. Observations:	833	AIC:	-348.7
Df Residuals:	829	BIC:	-329.8
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1722	0.019	8.849	0.000	0.134	0.210
age	0.3322	0.022	14.839	0.000	0.288	0.376
bmi	-0.1513	0.035	-4.310	0.000	-0.220	-0.082
children	0.0504	0.028	1.773	0.077	-0.005	0.106

Omnibus:	327.475	Durbin-Watson:	2.029
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1007.593
Skew:	1.992	Prob(JB):	1.60e-219
Kurtosis:	6.628	Cond. No.	6.85

const	8.23
age	1.02
bmi	1.02
children	1.00)

$$y = 0.1722 + 0.3322x_1 - 0.1513x_3 + 0.0504x_4$$

ANALISIS MODEL KETIGA

OLS Regression Results

Dep. Variable:	charges	R-squared:	0.595
Model:	OLS	Adj. R-squared:	0.594
Method:	Least Squares	F-statistic:	610.5
Date:	Tue, 19 Dec 2023	Prob (F-statistic):	9.12e-164
Time:	11:55:39	Log-Likelihood:	452.42
No. Observations:	833	AIC:	-898.8
Df Residuals:	830	BIC:	-884.7
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0549	0.009	6.187	0.000	0.038	0.072
age	0.3534	0.016	22.131	0.000	0.322	0.385
smoker	0.4298	0.015	28.554	0.000	0.400	0.459

Omnibus:	535.021	Durbin-Watson:	2.109
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3838.329
Skew:	3.035	Prob(JB):	0.00
Kurtosis:	11.588	Cond. No.	4.07

const	3.31
age	1.01
smoker	1.01)

$$y = 0.0549 + 0.3534x_1 + 0.4298x_5$$

ANALISIS MODEL KEEMPAT

OLS Regression Results

Dep. Variable:	charges	R-squared:	0.599
Model:	OLS	Adj. R-squared:	0.598
Method:	Least Squares	F-statistic:	413.5
Date:	Tue, 19 Dec 2023	Prob (F-statistic):	3.72e-164
Time:	11:55:44	Log-Likelihood:	456.67
No. Observations:	833	AIC:	-905.3
Df Residuals:	829	BIC:	-886.4
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0475	0.009	5.160	0.000	0.029	0.066
age	0.3699	0.017	21.917	0.000	0.337	0.403
smoker	0.4870	0.025	19.722	0.000	0.439	0.535
age_smoker	-0.1464	0.050	-2.916	0.004	-0.245	-0.048

Omnibus:	528.828	Durbin-Watson:	2.107
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3748.989
Skew:	2.992	Prob(JB):	0.00
Kurtosis:	11.497	Cond. No.	12.5

const	3.59
age_smoker	2.79
smoker	2.73
age	1.13)

$$y = 0.0475 + 0.3699x_1 + 0.4870x_5 - 0.1464x_1x_5$$

ANALISIS MODEL KELIMA

OLS Regression Results

Dep. Variable:	charges	R-squared:	0.603
Model:	OLS	Adj. R-squared:	0.602
Method:	Least Squares	F-statistic:	420.1
Date:	Tue, 19 Dec 2023	Prob (F-statistic):	7.14e-166
Time:	11:55:48	Log-Likelihood:	460.64
No. Observations:	833	AIC:	-913.3
Df Residuals:	829	BIC:	-894.4
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0844	0.011	7.405	0.000	0.062	0.107
age	0.1276	0.058	2.211	0.027	0.014	0.241
smoker	0.4297	0.015	28.818	0.000	0.400	0.459
age_kuadrat	0.2431	0.060	4.066	0.000	0.126	0.360

Omnibus:	536.856	Durbin-Watson:	2.095
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3888.231
Skew:	3.044	Prob(JB):	0.00
Kurtosis:	11.658	Cond. No.	19.8

age	13.38
age_kuadrat	13.38
const	5.56
smoker	1.01)

$$y = 0.0844 + 0.1276x_1 + 0.4297x_5 + 0.2431x_1^2$$

ANALISIS MODEL KEENAM

OLS Regression Results

Dep. Variable:	charges	R-squared:	0.607
Model:	OLS	Adj. R-squared:	0.605
Method:	Least Squares	F-statistic:	319.5
Date:	Tue, 19 Dec 2023	Prob (F-statistic):	3.46e-166
Time:	11:55:54	Log-Likelihood:	464.50
No. Observations:	833	AIC:	-919.0
Df Residuals:	828	BIC:	-895.4
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0766	0.012	6.541	0.000	0.054	0.100
age	0.1496	0.058	2.577	0.010	0.036	0.264
smoker	0.4838	0.024	19.754	0.000	0.436	0.532
age_smoker	-0.1382	0.050	-2.775	0.006	-0.236	-0.040
age_kuadrat	0.2363	0.060	3.965	0.000	0.119	0.353

Omnibus:	531.381	Durbin-Watson:	2.096
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3820.974
Skew:	3.004	Prob(JB):	0.00
Kurtosis:	11.601	Cond. No.	19.9

age	13.64
age_kuadrat	13.40
const	5.91
age_smoker	2.79
smoker	2.73)

$$y = 0.0766 + 0.1496x_1 + 0.4838x_5 - 0.1382x_5x_1 + 0.2363x_1^2$$

ANALISIS MODEL KETUJUH

OLS Regression Results

Dep. Variable:	charges	R-squared:	0.618
Model:	OLS	Adj. R-squared:	0.615
Method:	Least Squares	F-statistic:	222.8
Date:	Tue, 19 Dec 2023	Prob (F-statistic):	7.30e-169
Time:	11:55:58	Log-Likelihood:	476.57
No. Observations:	833	AIC:	-939.1
Df Residuals:	826	BIC:	-906.1
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0545	0.016	3.380	0.001	0.023	0.086
age	0.0534	0.061	0.880	0.379	-0.066	0.173
smoker	0.4905	0.024	20.145	0.000	0.443	0.538
age_smoker	-0.1404	0.049	-2.851	0.004	-0.237	-0.044
age_kuadrat	0.3335	0.062	5.337	0.000	0.211	0.456
bmi	0.0316	0.026	1.239	0.216	-0.018	0.082
children	0.1012	0.021	4.780	0.000	0.060	0.143

Omnibus:	541.056	Durbin-Watson:	2.096
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4001.358
Skew:	3.065	Prob(JB):	0.00
Kurtosis:	11.815	Cond. No.	23.1

age	15.32
age_kuadrat	15.13
const	11.52
age_smoker	2.80
smoker	2.77
children	1.13
bmi	1.10)

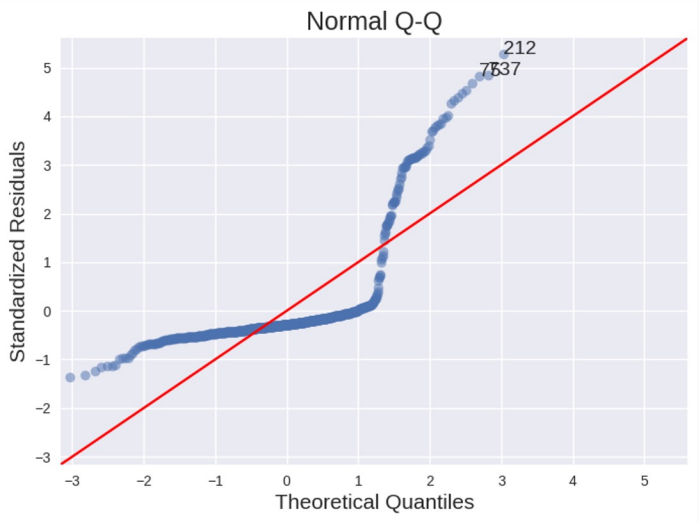
$$y = 0.0545 + 0.0534x_1 + 0.4905x_5 - 0.1404x_5x_1 + 0.3335x_1^2 + 0.0316x_3 + 0.1012x_4$$

OUTPUT, ANALISIS, DAN ASUMSI MODEL FINAL

Model final: model keempat

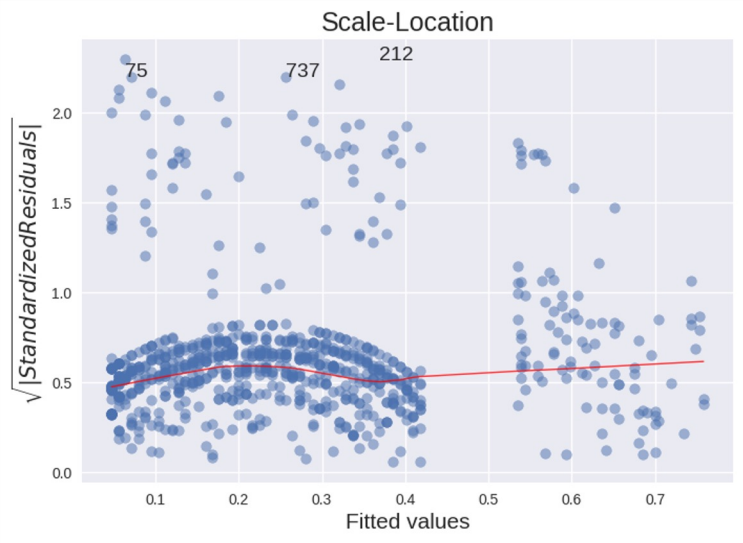
```
          const   3.59
age_smoker        2.79
        smoker    2.73
           age    1.13)
```

VIF

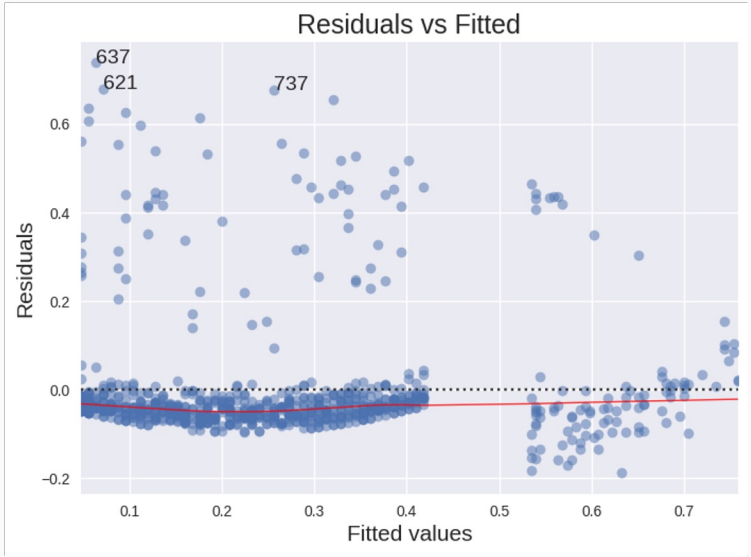


Normality Test

OUTPUT, ANALISIS, DAN ASUMSI MODEL FINAL



Homoscedasticity Test



Linearity Test

INTERPRETASI MODEL AKHIR

$$y = 0.0475 + 0.3699x_1 + 0.4870x_5 - 0.1464x_1x_5$$

Interpretasi :

1. Intercept (Intersepsi): Intercept, yaitu 0.0475, adalah nilai y (biaya asuransi) ketika semua variabel independen (x_1, x_5) adalah nol. Ini mewakili nilai y (biaya asuransi) ketika tidak ada kontribusi dari variabel independen.
2. Koefisien x_1 (Usia) : Koefisien 0.3699 untuk x_1 menunjukkan bahwa setiap peningkatan satu unit pada x_1 akan menghasilkan peningkatan sebesar 0.3699 pada nilai y (biaya asuransi), dengan asumsi bahwa variabel lainnya tetap konstan.
3. Koefisien x_5 (Perokok atau bukan) : Koefisien 0.4870 untuk x_5 menandakan bahwa setiap peningkatan satu unit pada x_5 akan menghasilkan peningkatan sebesar 0.4870 pada nilai y (biaya asuransi), dengan asumsi variabel lainnya tetap konstan.
4. Koefisien x_1x_5 : Koefisien -0.1464 adalah koefisien interaksi antara x_1 (Usia) dan x_5 (Perokok atau bukan).

TERIMA KASIH