FInal Project Model Linier

Kelompok 8

Anggota Kelompok

- Muhammad Rayhan N. A. - 2006571053
- Rifqi Hafizuddin - 2106638204
- Zahrah Aulia Putri - 2106724896
- Zahrah Mahfuzah - 2106704004

Pendahuluan

Premi asuransi adalah jumlah uang yang harus dibayarkan secara reguler oleh seseorang atau institusi untuk menikmati polis asuransi. Berdasarkan definisi tersebut, regresi linear berganda akan digunakan untuk mengetahui hubungan antara peubah terikat dan peubah bebas yang bertujuan untuk memprediksi biaya pengobatan dari suatu individu kedepannya dan membantu asuransi kesehatan dalam memutuskan besarnya premi yang harus mereka kenakan terhadap peserta asuransi.

Nama Peubah	Definisi	Skala Pengukuran	Tipe Data	Faktor
Expenses (Y)	Total biaya pengobatan peserta asuransi	Rasio	Numerik	-
Age (X1)	Usia peserta asuransi	Rasio	Numerik	-
Sex (X2)	Jenis kelamin dari peserta asuransi	Nominal	Kategorik	Wanita (Female) Pria (Male)
Bmi (X3)	Indeks massa tubuh peserta asuransi	Rasio	Numerik	-
Children (X4)	Jumlah anak dari peserta asuransi	Rasio	Numerik	-
Smoker (X5)	Apakah peserta asuransi tersebut merupakan perokok atau bukan	Nominal	Kategorik	Perokok (Yes) Non Perokok (No)
Region (X6)	Daerah peserta asuransi tinggal	Nominal	Kategorik	Tenggara (southeast) Barat daya (southwest) Timur laut (northeast) Barat laut (northwest)

Pre-Processing

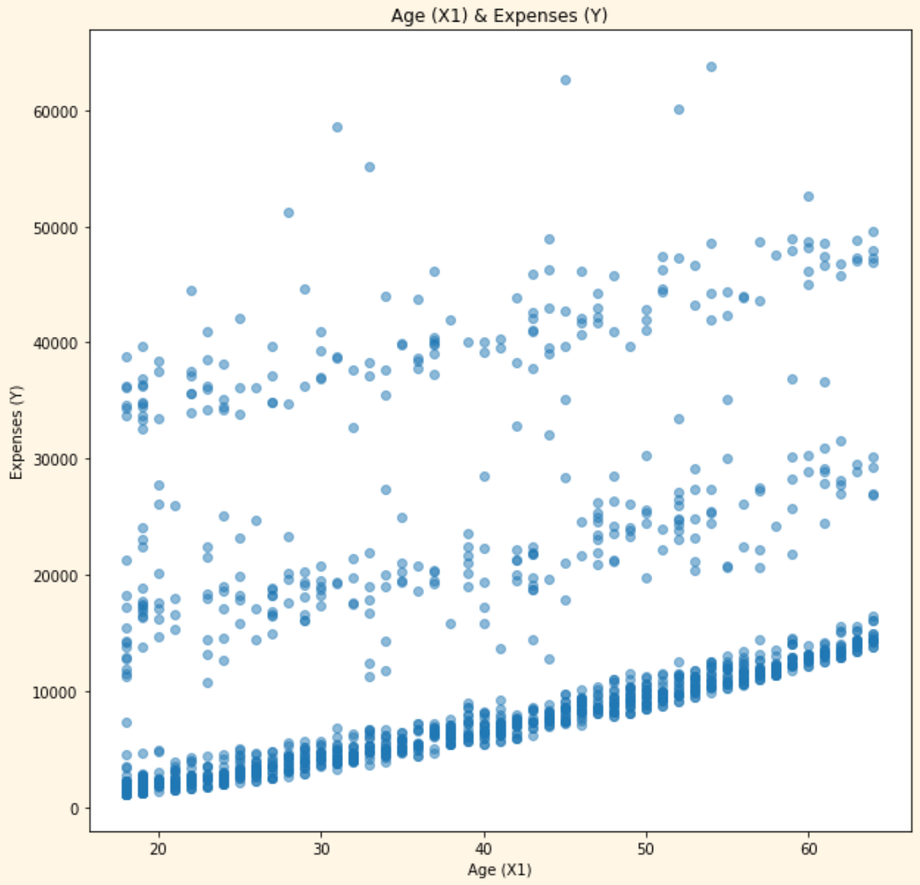
Pada tahap pre-processing, akan dilakukan importing terhadap data terlebih dahulu untuk memasukkan data ke software pengolahan data. Kemudian, dilakukan pengecekan struktur data dan tahap-tahap pre-processing lainnya untuk mengatasi outlier pada data, menyeleksi variabel-variabel yang dibutuhkan, dan melihat hubungan atau keterkaitan antar variabel.

Langkah-langkah yang dilakukan dalam pre-processing:

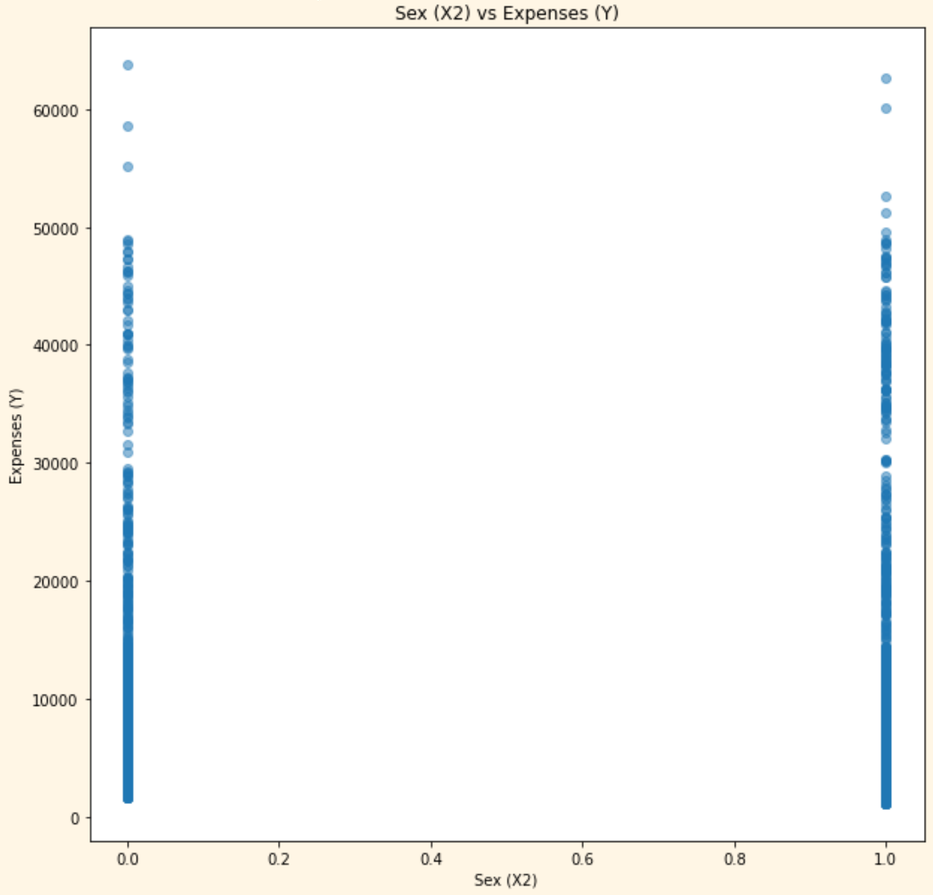
1. Mendeteksi adanya missing values: Dihasilkan bahwa tidak terdapat missing values pada data
2. Mendeteksi adanya duplicates pada data: Diperoleh bahwa terdapat 1 duplicates pada data
3. Drop duplicates yang terdapat pada data

Selanjutnya, dilakukan analisis deskriptif dengan visualisasi untuk melihat hubungan setiap variabel prediktor terhadap variabel respons.

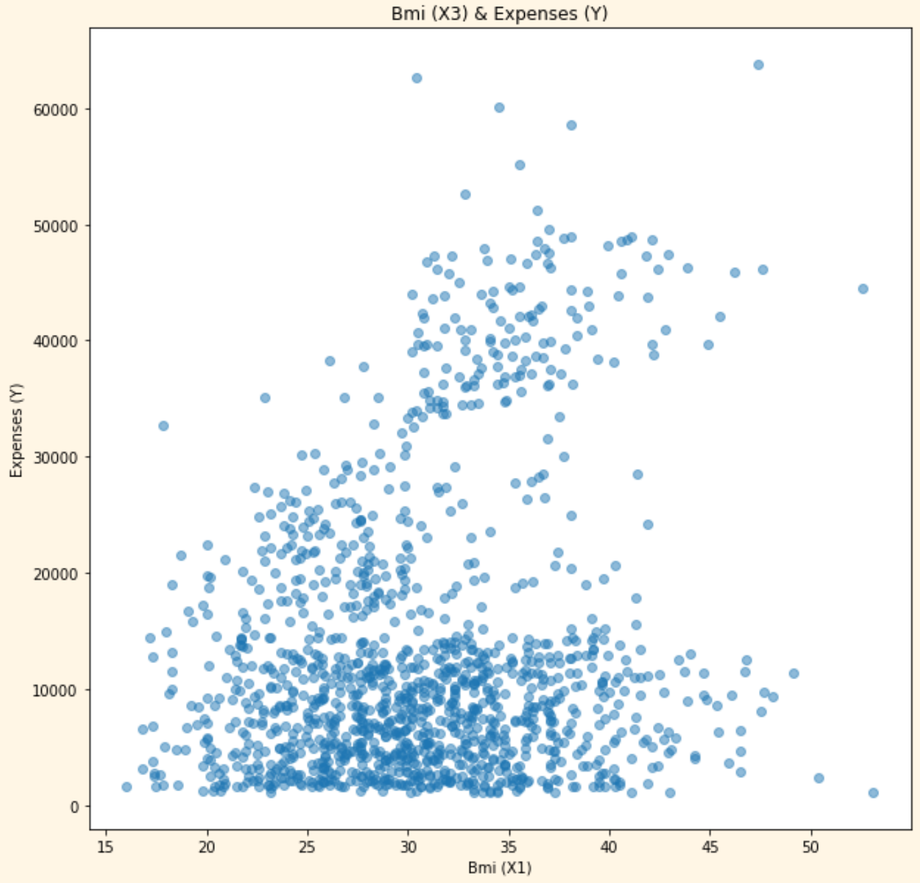
1). Age (X1) dan Expenses (Y)



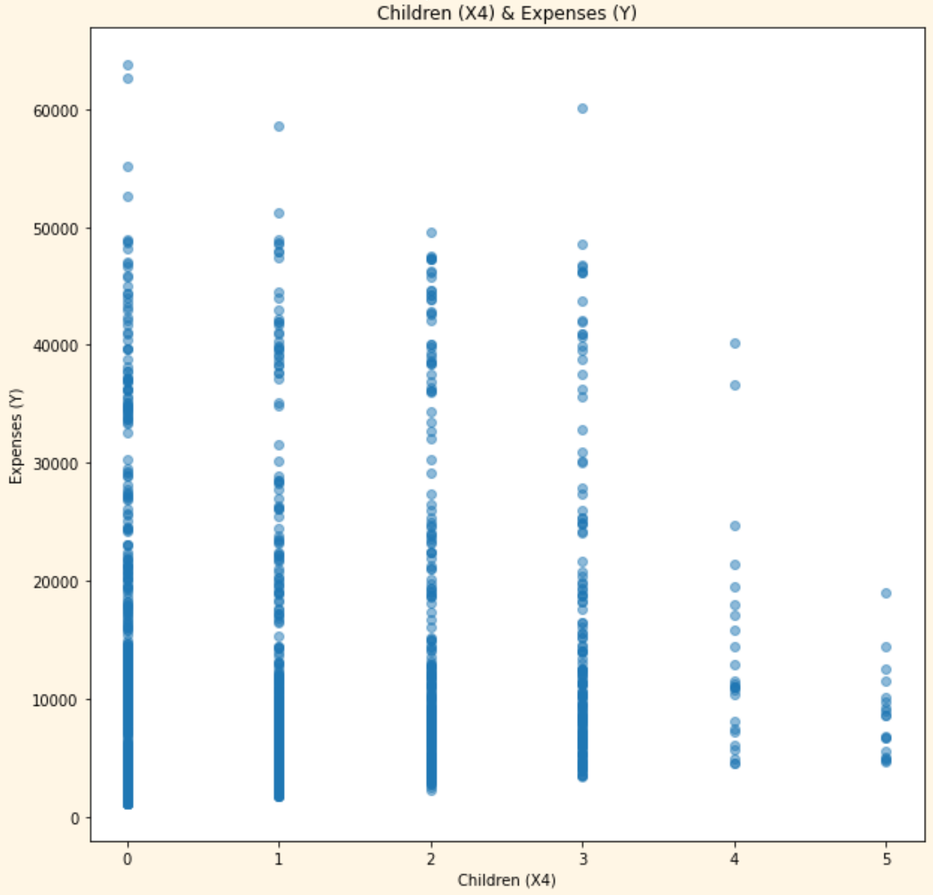
2). Sex (X2) dan Expenses (Y)



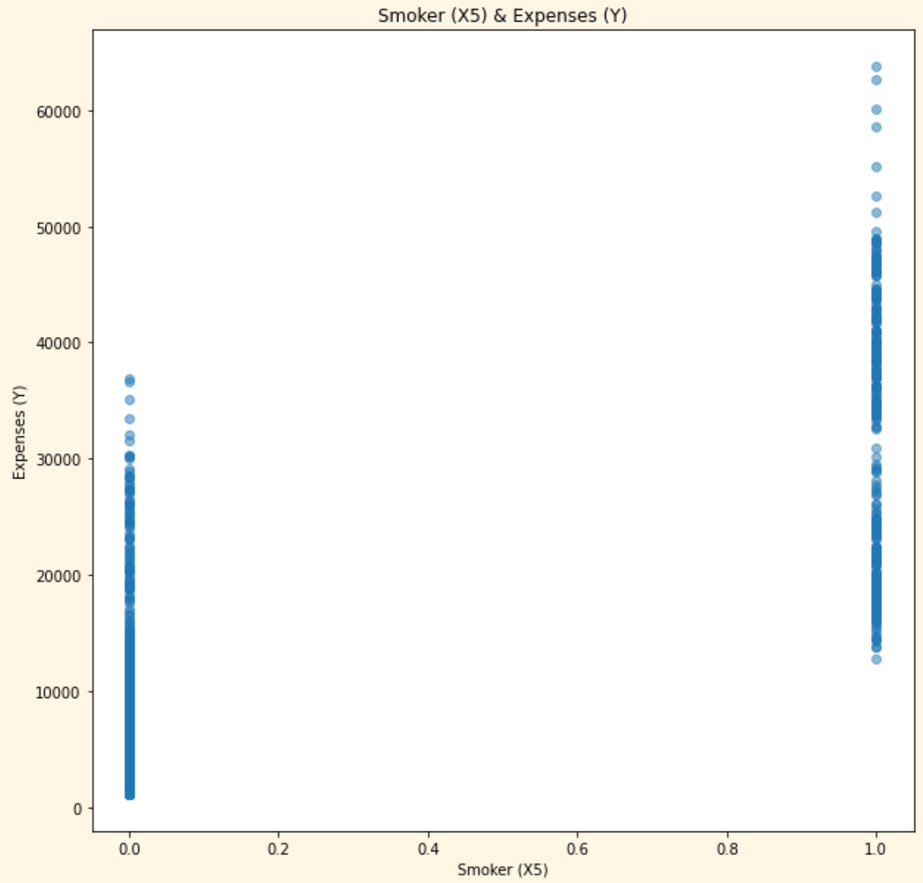
3). Bmi (X3) dan Expenses (Y)



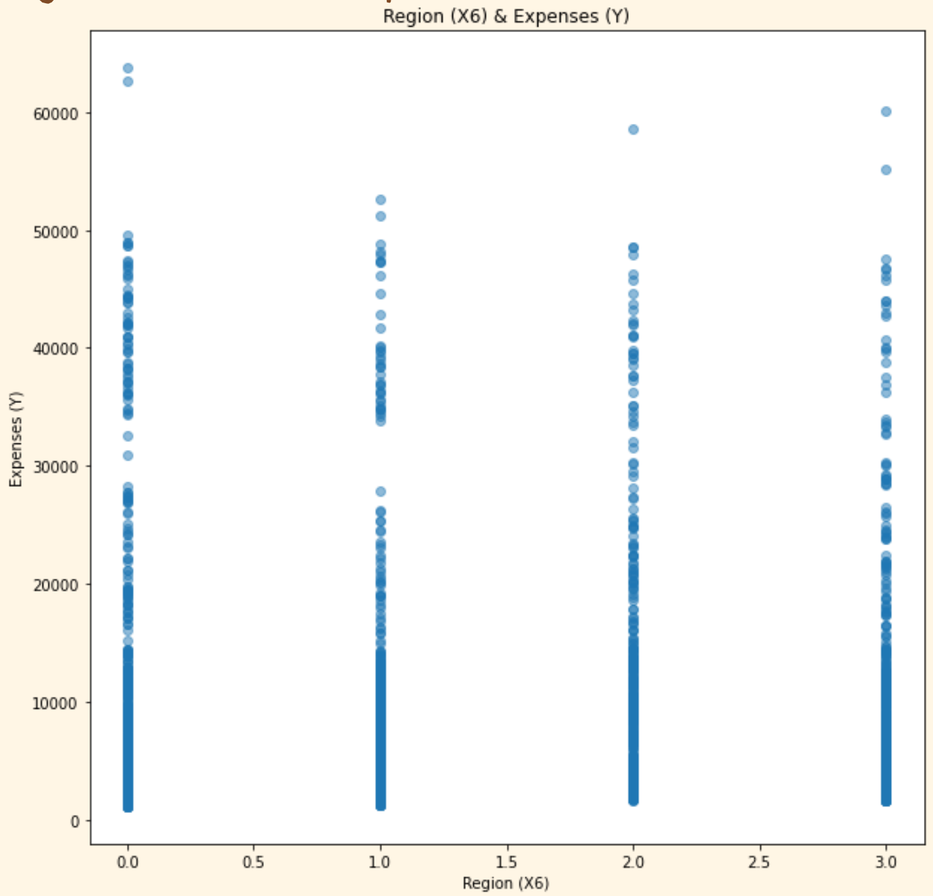
4). Children (X4) dan Expenses (Y)



5). Smoker (X5) dan Expenses (Y)



6). Region (X6) dan Expenses (Y)



Kemudian, akan dicari ringkasan statistik dari dataset sebagai berikut :

	age	sex	bmi	children	smoker	region	expenses
count	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	0.505232	30.665471	1.094918	0.204783	1.455904	13270.422414
std	14.049960	0.500160	6.098382	1.205493	0.403694	1.130888	12110.011240
min	18.000000	0.000000	16.000000	0.000000	0.000000	0.000000	1121.870000
25%	27.000000	0.000000	26.300000	0.000000	0.000000	0.000000	4740.287500
50%	39.000000	1.000000	30.400000	1.000000	0.000000	1.000000	9382.030000
75%	51.000000	1.000000	34.700000	2.000000	0.000000	2.000000	16639.915000
max	64.000000	1.000000	53.100000	5.000000	1.000000	3.000000	63770.430000

Pemodelan

Model yang diajukan:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon,\ \epsilon \sim NIID(0, \sigma^2)$$

Keterangan:

Y = *expenses*

β_0 = *intercept*

β_i = koefisien regresi yang menunjukkan pengaruh variabel X_i terhadap variabel Y.

X_1 = *age*

X_2 = bmi

X_3 = *children*

X_4 = *smoker1*, *dummy variable* dari variabel kategorik *smoker* yang memiliki dua level, bernilai 1 untuk *yes*, 0 untuk *no*.

$X_5 = X_2 X_4$ = interaksi antara variabel bmi dengan variabel *smoker*

ϵ = error, menunjukkan pengaruh yang tidak teramati terhadap variabel Y.

Berdasarkan estimasi parameter dari R, model menjadi

$$Y = -2289,23 + 265,70X_1 - 6,45X_2 + 491,93X_3 - 20415,21X_4 + 1436,34X_5 + \epsilon$$

Asumsi yang harus dipenuhi model, yaitu:

1. Linearitas
2. Normalitas
3. Homoskedastisitas
4. Nonmultikolinearitas

Tahapan Pembentukan Model

1 Membentuk model dari metode regresi stepwise

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

Keterangan:

Y = *expenses*

β_0 = *intercept*

β_i = koefisien regresi yang menunjukkan pengaruh variabel X_i terhadap variabel Y.

X_1 = *age*

X_2 = bmi

X_3 = *children*

X_4 = *smoker1*, *dummy variable* dari variabel kategorik *smoker* yang memiliki dua level, bernilai 1 untuk *yes*, 0 untuk *no*.

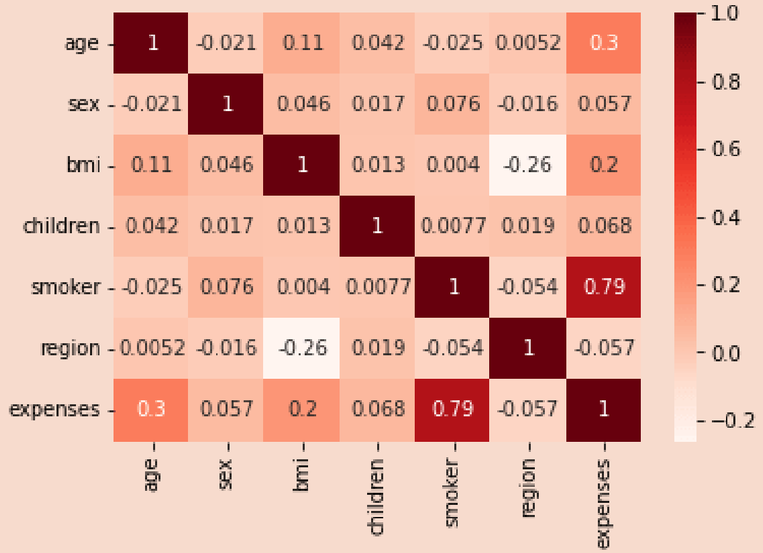
ϵ = error, menunjukkan pengaruh yang tidak teramati terhadap variabel Y.

```
> summary(stepwise)

Call:
lm(formula = expenses ~ smoker + age + bmi + children, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-12156.2  -2783.5   -947.3   1311.9  26229.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11584.08    1031.13 -11.234  < 2e-16 ***
smoker1      23928.24     454.76  52.617  < 2e-16 ***
age            256.06      13.00  19.697  < 2e-16 ***
bmi            305.83      29.85  10.245  < 2e-16 ***
children       407.39     151.94   2.681  0.00745 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5978 on 1065 degrees of freedom
Multiple R-squared:  0.7535,    Adjusted R-squared:  0.7526
F-statistic: 814.1 on 4 and 1065 DF,  p-value: < 2.2e-16
```

Tahapan Pembentukan Model

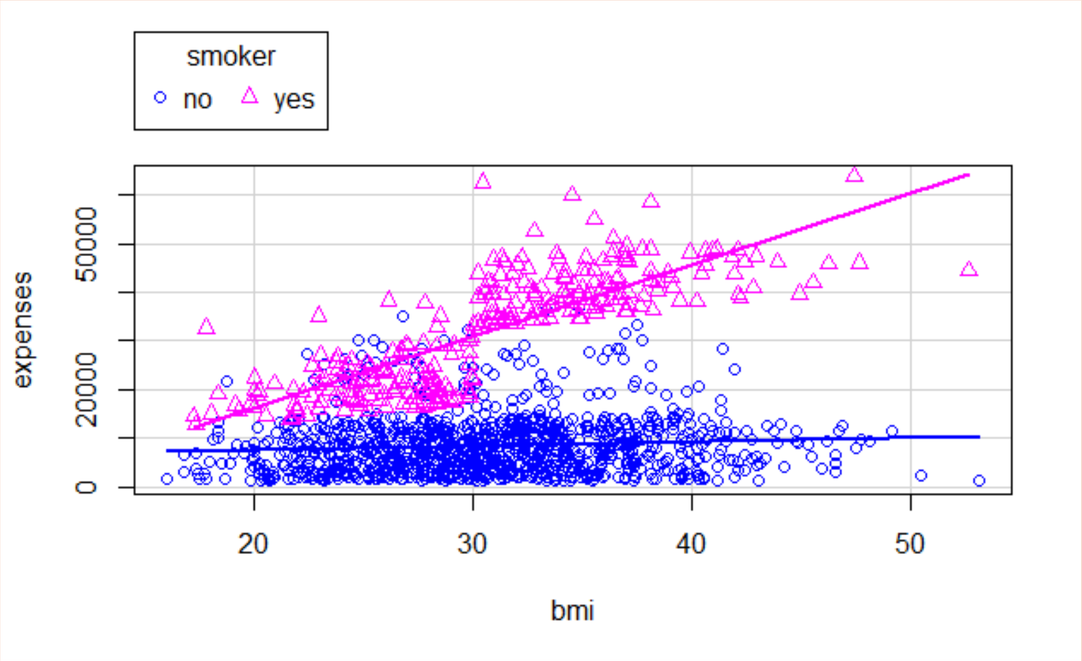
2 Membentuk model 2

Dari heatmap di bawah ini, dapat dilihat bahwa korelasi antara variabel children dengan variabel expenses cukup kecil, yaitu sekitar 0.068, sehingga kami mencoba untuk menghilangkan variabel children dari model, walaupun pada model sebelumnya variabel chidlren signifikan.



Tahapan Pembentukan Model

Kemudian, dibentuk plot antara variabel bmi dengan variabel expenses yang dikelompokkan berdasarkan variabel smoker. Dapat dilihat dari plot, semakin besar nilai bmi, maka nilai expenses akan besar jika orang tersebut adalah perokok, tetapi nilai expenses cenderung tetap jika orang tersebut bukan perokok. Hal ini mengindikasikan terdapat interaksi antara variabel bmi dengan variabel smoker.



Tahapan Pembentukan Model

Dibentuk model modifikasi dari hasil metode regresi stepwise yang mengandung interaksi antara variabel bmi dengan variabel smoker.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

Keterangan:

Y = *expenses*

β_0 = *intercept*

β_i = koefisien regresi yang menunjukkan pengaruh variabel X_i terhadap variabel Y.

X_1 = *age*

X_2 = bmi

X_3 = *smoker1*, *dummy variable* dari variabel kategorik *smoker* yang memiliki dua level, bernilai 1 untuk *yes*, 0 untuk *no*.

$X_4 = X_2 X_3$ = interaksi antara variabel bmi dengan variabel *smoker*

ϵ = error, menunjukkan pengaruh yang tidak teramati terhadap variabel Y.

```
> summary(model2)

Call:
lm(formula = expenses ~ age + bmi + smoker + bmi * smoker, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-14669.0  -1918.4  -1255.3   -250.4  24494.8

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2015.648    896.071  -2.249   0.0247 *
age            264.283     10.369  25.487   <2e-16 ***
bmi             -1.149     26.844  -0.043   0.9659
smoker1     -20300.022   1814.627 -11.187   <2e-16 ***
bmi:smoker1   1439.778     57.876  24.877   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4770 on 1065 degrees of freedom
Multiple R-squared:  0.8431,    Adjusted R-squared:  0.8425
F-statistic:  1430 on 4 and 1065 DF,  p-value: < 2.2e-16
```

Tahapan Pembentukan Model

3 Membentuk model 3

Variabel children akan dicoba untuk kembali diikutsertakan dalam model dengan tujuan untuk memperbesar nilai adjusted r-squared.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$$

Keterangan:

Y = *expenses*

β_0 = *intercept*

β_i = koefisien regresi yang menunjukkan pengaruh variabel X_i terhadap variabel Y.

X_1 = *age*

X_2 = bmi

X_3 = *children*

X_4 = *smoker1*, *dummy variable* dari variabel kategorik *smoker* yang memiliki dua level, bernilai 1 untuk *yes*, 0 untuk *no*.

$X_5 = X_2 X_4$ = interaksi antara variabel bmi dengan variabel *smoker*

ϵ = error, menunjukkan pengaruh yang tidak teramati terhadap variabel Y.

```
> summary(model3)

Call:
lm(formula = expenses ~ age + bmi + children + smoker + bmi *
    smoker, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-14751.7  -1830.8  -1249.6   -375.5  24520.0

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2406.371    894.157  -2.691  0.00723 **
age            262.639     10.298  25.504  < 2e-16 ***
bmi             -4.095     26.649  -0.154  0.87791
children       502.423    120.382   4.174 3.24e-05 ***
smoker1     -20533.325   1801.667 -11.397  < 2e-16 ***
bmi:smoker1   1447.295     57.464  25.186  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4734 on 1064 degrees of freedom
Multiple R-squared:  0.8456,    Adjusted R-squared:  0.8449
F-statistic:  1165 on 5 and 1064 DF,  p-value: < 2.2e-16
```

Pengolahan Data dan Analisis Hasil

1. Kriteria Model Terbaik

Terdapat beberapa kriteria yang harus dipenuhi untuk menentukan model terbaik. Berikut kriteria yang sebaiknya ada dalam sebuah model :

1. Model memenuhi asumsi normalitas, homoskedastisitas, linearitas, dan non multikolinearitas.
2. Model memberikan hasil penolakan hipotesis null pada uji F dan uji T untuk setiap variabel prediktor.
3. Model memenuhi prinsip parsimony dimana model berbentuk sederhana dengan parameter lebih sedikit akan lebih disukai dibanding model yang kompleks dengan parameter lebih banyak.
4. Model memenuhi All-Possible-Regressions Selection Procedure

2. Asumsi Model

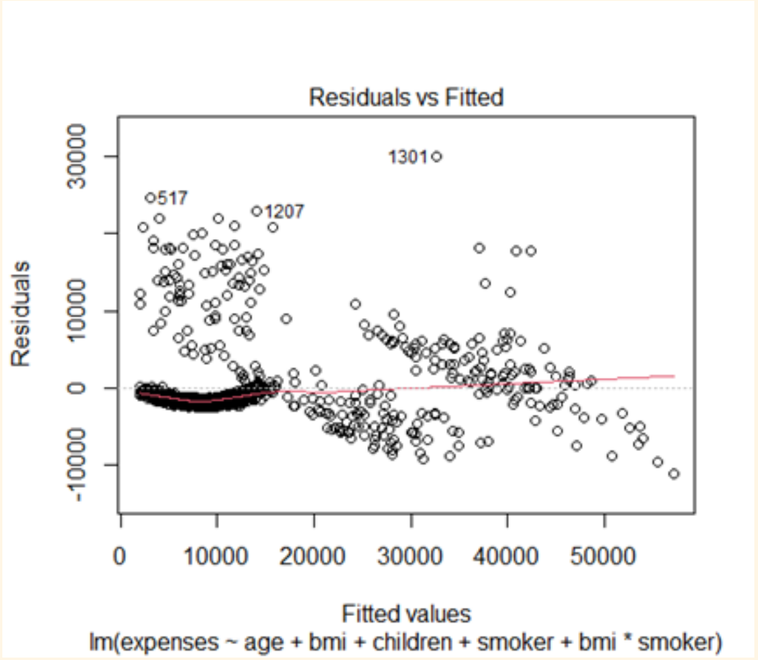
1) Linieritas

Hipotesis:

H0 : Model regresi bersifat linier (p-value >α)

H1 : Model regresi bersifat nonlinier (p-value ⩽α)

```
> lmtest::resettest(model3, power = 2)

        RESET test

data:  model3
RESET = 1.5036, df1 = 1, df2 = 1063, p-value =
0.2204
```



Tidak terdapat pola tertentu pada grafik tersebut. dan p-value dari Uji Reset lebih besar dari α =0.05 sehingga H0 tidak ditolak dan asumsi linearitas terpenuhi.

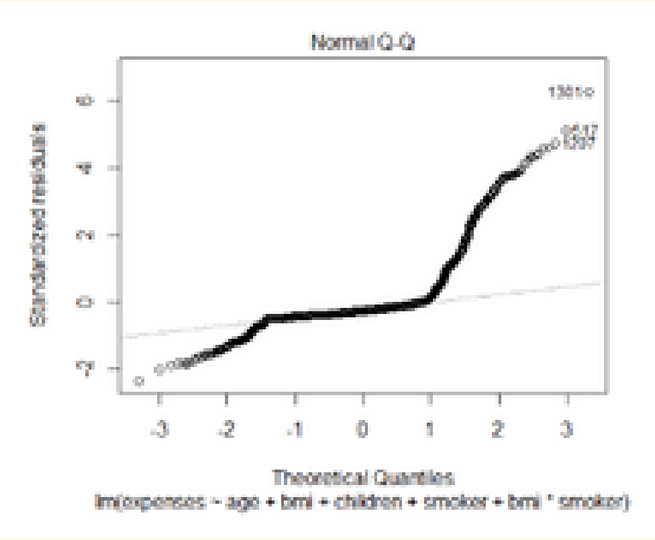
2) Normalitas Residual

Hipotesis:

H0 : Data menyebar secara normal (p-value >α)

H1 : Data menyebar secara tidak normal (p-value ⩽α)

```
> shapiro.test(train$expenses)

        Shapiro-Wilk normality test

data:  train$expenses
W = 0.81472, p-value < 2.2e-16
```



Berdasarkan p-value dari uji Shapiro-Wilk, H0 ditolak sehingga dengan α =0.05, dapat dinyatakan bahwa data menyebar secara tidak normal. Berdasarkan grafik normal q-q diatas, dapat dikatakan bahwa titik tidak berada cukup dekat dengan garis sehingga asumsi error berdistribusi normal dengan mean 0 tidak terpenuhi.

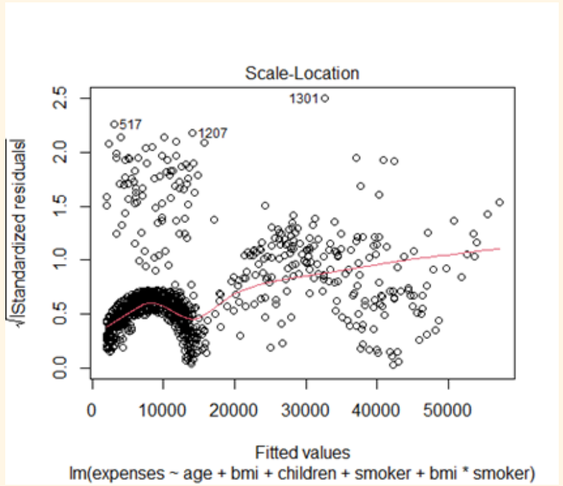
3) Homoskedastisitas

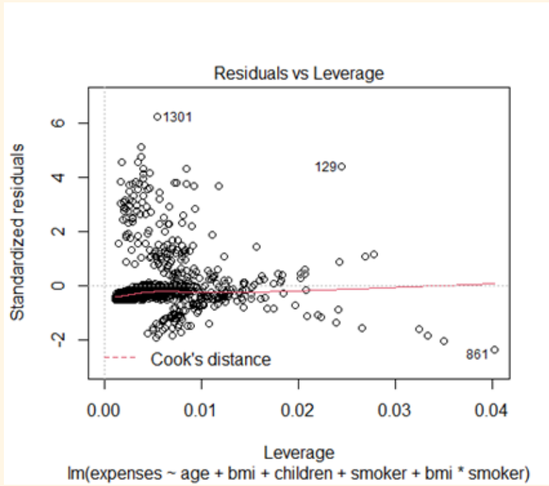
Hipotesis:

H0 : Ragam galat bersifat homoskedastisitas (p-value >α)

H1 : Ragam galat tidak bersifat heteroskedastisitas. (p-value ⩽α)

```
> lmtest::bptest(model3)

        studentized Breusch-Pagan test

data:  model3
BP = 7.3081, df = 5, p-value = 0.1987
```





Pada grafik scale-location, titik data tersebar cukup hampir sama di antara garis merah dan berdasarkan Uji bptest, dengan α =0.05, H0 tidak ditolak dan dapat dinyatakan bahwa asumsi homoskedastisitas terpenuhi.

4) Multikolinieritas

```
> vif(model3, type="predictor")
GVIFs computed for predictors
             GVIF Df GVIF^(1/(2*Df)) Interacts With      Other Predictors
age      1.010297  1        1.005135              --   bmi, children, smoker
bmi      1.007827  3        1.001300          smoker      age, children
children 1.003706  1        1.001851              --   age, bmi, smoker
smoker   1.007827  3        1.001300             bmi      age, children
```

Karena terdapat interaksi antar variabel. maka akan digunakan Generalized Variance Inflation Factor (GVIF). Nilai GVIF dari semua variabel lebih kecil dari 5 sehingga dapat disimpulkan tidak terdapat multikolinearitas variabel prediktor pada model.

3. Prosedur Pemilihan Model Regresi Terbaik

	p+1	Adj R-Squared	Adj R-Squared dari Test Data	AIC	PRESS	Cp
1	5	0.7526218	0.7338777	21652.73	38500485202	5.997952
2	5	0.8424826	0.8085263	21169.75	24502382934	-381.228840
3	6	0.8448741	0.8106489	21154.38	24148829937	-390.162004

Tabel Kriteria pemilihan model terbaik.

1. Adjusted R-Squared

$$R_{\mathrm{a}}^2 = 1 - (n-1)\left[\frac{\mathrm{MSE}}{\mathrm{SS(Total)}}\right]$$

R-squared merupakan rumus pembagian antara Sum Squared Regression dengan Sum Squared Total. tetapi nilainya akan meningkat dengan bertambahnya variabel. Maka akan digunakan Adjusted R-Squared yang sudah mempertimbangkan jumlah variabel pada model. Semakin besar nilai Adjusted R-Squared. maka model akan semakin baik. Dari tabel. model 3 memiliki nilai Adjusted R-Squared terbaik. yaitu 0.8447841.

2. AIC

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

Akaike's information criterion (AIC) merupakan metode analisis yang digunakan untuk memperoleh model faktor produksi yang terbaik dengan menggunakan estimasi maximum likelihood sebagai perhitungan yang sesuai. Semakin kecil nilai AIC. maka model akan semakin baik. Dari tabel. model 3 memiliki nilai AIC terbaik. yaitu 21154.38.

3. PRESS

$$\text{PRESS} = \sum_{i=1}^{n} [y_i - \hat{y}_{(i)}]^2$$

Prediction Sum of Squares (PRESS) merupakan metode yang menghitung kuadrat selisih dari nilai yi dengan estimasi yi. Karena selisih yang kecil mengindikasikan model cukup akurat, maka model yang baik memiliki nilai PRESS yang kecil. Dari tabel, model 3 memiliki nilai PRESS terbaik, yaitu 24148829937.

4. Cp

$$C_p = \frac{\text{SSE}_p}{\text{MSE}_k} + 2(p+1) - n$$

Cp adalah nilai statistik yang fokus pada meminimalisir jumlah mean square error dan bias regresi, sehingga akan dipilih model dengan nilai cp terkecil dan mendekati p+1, yaitu model 1 dengan nilai cp = 5.997952. Namun, jika kita hanya ingin fokus dengan meminimalisir jumlah mean square error, maka akan dipilih model dengan nilai cp terkecil, yaitu model 3 dengan nilai cp = -390.162004.

4. Uji signifikansi Parameter

4.4. Uji signifikansi Parameter

Kami akan menguji apakah terdapat pengaruh variabel prediktor pada model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$, dengan X_5 merupakan interaksi antara variabel bmi dengan variabel *smoker,* terhadap variabel responnya menggunakan uji F.

Hipotesis

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

H_1 : setidaknya terdapat satu $\beta_j \neq 0$, $j = 1, 2, \ldots, 5$

Tingkat Signifikansi

Tingkat signifikansi yang digunakan adalah 5%

Daerah Penolakan

H_0 ditolak jika $F > F_{\alpha,p,n-(p+1)}$ atau $p\text{-}value < \alpha = 0.05$

4. Uji signifikansi Parameter

Statistik Uji

Berdasarkan hasil summary dari model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon$

```
> model3 <- lm(expenses ~ age+bmi+children+smoker+bmi*smoker,data=train)
> summary(model3)

Call:
lm(formula = expenses ~ age + bmi + children + smoker + bmi *
    smoker, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-14751.7  -1830.8  -1249.6   -375.5  24520.0

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2406.371    894.157  -2.691  0.00723 **
age            262.639     10.298  25.504  < 2e-16 ***
bmi             -4.095     26.649  -0.154  0.87791
children       502.423    120.382   4.174 3.24e-05 ***
smoker1     -20533.325   1801.667 -11.397  < 2e-16 ***
bmi:smoker1   1447.295     57.464  25.186  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4734 on 1064 degrees of freedom
Multiple R-squared:  0.8456,    Adjusted R-squared:  0.8449
F-statistic:  1165 on 5 and 1064 DF,  p-value: < 2.2e-16
```

4. Uji signifikansi Parameter

Keputusan

Dapat dilihat p-value untuk uji F sangatlah kecil, yaitu sebesar < 2.2 x 10^-16. Dengan tingkat signifikansi α =0.05, H0 ditolak. Jadi, terdapat variabel prediktor yang dapat menjelaskan variabel respon.

Kesimpulan

Berdasarkan hasil analisis yang kami dapatkan, model 3 dipilih sebagai model terbaik dengan nilai Adjusted R-Squared sebesar 0.8456 atau sekitar 84.5% peubah terikat Y (expenses) dapat dijelaskan oleh peubah bebas yang dimana peubah bebas yang signifikan adalah umur (X1), jumlah anak (X3), perokok atau bukan (X4), dan interaksi antara BMI dan smoker (X5), akan tetapi untuk peubah bebas BMI (X2) tidaklah signifikan dikarenakan p-value-nya yang terlalu besar.

Thank You