INSURANCE CHARGES

BAGIAN 1
PENDAHULUAN

1.1 Latar Belakang

Dalam kehidupan sehari-hari, asuransi kesehatan menjadi salah satu perlindungan finansial yang penting untuk mengatasi biaya medis yang tinggi. Untuk perusahaan asuransi, penentuan premi yang adil dan akurat merupakan hal yang sangat krusial agar tetap menjaga keseimbangan antara keuntungan perusahaan dan kepuasan pelanggan. Dalam proses penetapan premi, terdapat berbagai faktor yang mempengaruhi besar kecilnya biaya yang harus dibayar oleh individu, seperti usia, jenis kelamin, status merokok, indeks massa tubuh (BMI), jumlah anak tanggungan, dan lokasi geografis. Mengetahui hubungan antar faktor-faktor tersebut dapat membantu perusahaan asuransi dalam merumuskan kebijakan premi yang lebih tepat sasaran.
Berdasarkan hal tersebut, kami tertarik untuk meneliti hubungan antara faktor-faktor yang mempengaruhi biaya asuransi kesehatan dengan menggunakan regresi berganda untuk memprediksi biaya asuransi yang harus dibayar oleh individu. Dengan membangun model regresi yang tepat, diharapkan perusahaan asuransi dapat lebih mudah menentukan premi berdasarkan karakteristik individu dan faktor-faktor lain yang relevan.

Next Page

BAGIAN 1
PENDAHULUAN

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, rumusan masalah dari penelitian ini adalah sebagai berikut:

1. Bagaimana model regresi terbaik untuk memprediksi biaya asuransi kesehatan yang harus dibayar oleh individu?
2. Apa saja faktor-faktor yang memiliki pengaruh signifikan terhadap biaya asuransi kesehatan berdasarkan model regresi?

1.3 Tujuan

Tujuan dari penelitian ini adalah sebagai berikut:

1. Membentuk model regresi terbaik dalam memprediksi biaya asuransi kesehatan yang harus dibayar oleh individu.
2. Menentukan variabel-variabel yang mempunyai pengaruh signifikan terhadap biaya asuransi kesehatan berdasarkan model regresi.

Next Page

BAGIAN 1
PENDAHULUAN

1.4 Pendefinisian Variabel

Dataset yang digunakan dalam penelitian ini dapat diakses pada link:https://www.kaggle.com/datasets/mirichoi0218/insurance

Data ini berisi informasi mengenai individu yang terdaftar dalam asuransi kesehatan, dengan variabel-variabel sebagai berikut:

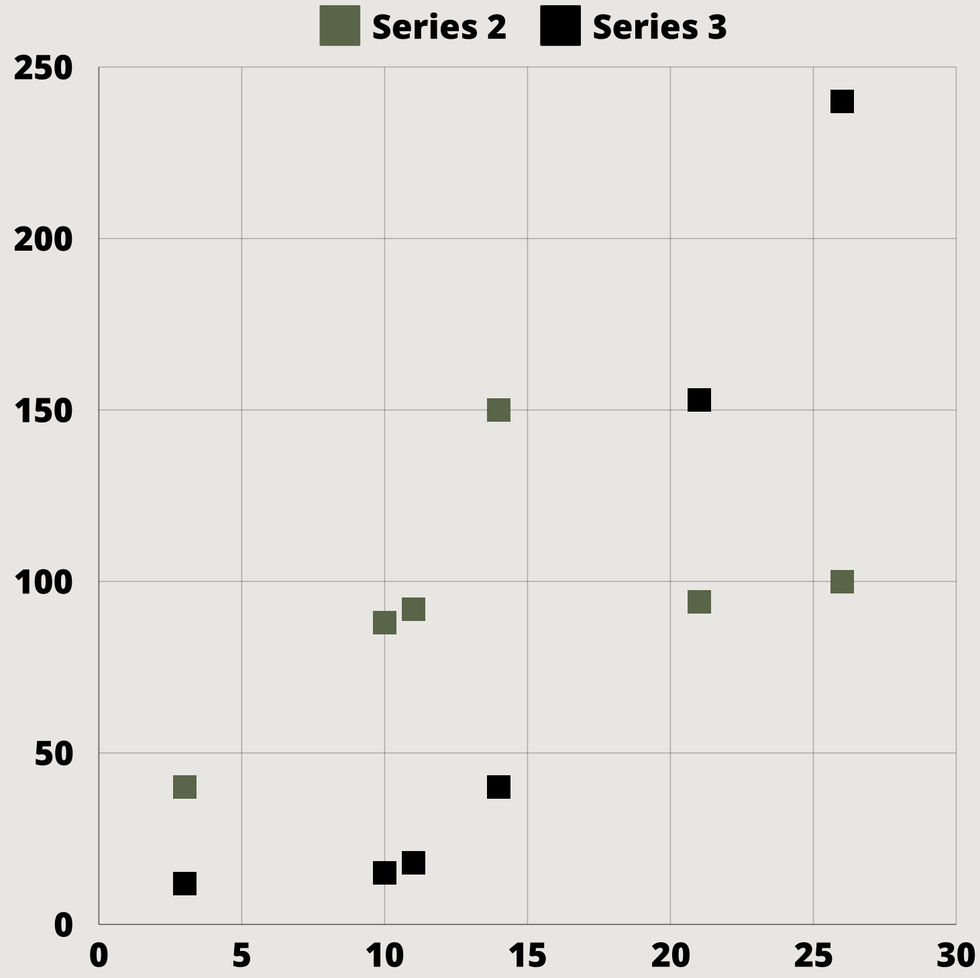
- Age: Usia individu
- Sex: Jenis kelamin individu (male/female)
- BMI: Indeks massa tubuh individu
- Children: Jumlah anak tanggungan individu
- Smoker: Status merokok individu (yes/no)
- Region: Lokasi geografis individu (Northwest, Northeast, Southeast, Southwest)
- Charges: Biaya asuransi yang dibayar oleh individu (variabel dependen)

Data ini mencakup informasi dari berbagai individu dengan karakteristik yang berbeda-beda, yang dapat digunakan untuk membangun model regresi guna memprediksi biaya asuransi berdasarkan faktor-faktor yang ada.

Next Page

UNIVERSITAS INDONESIA

BAGIAN 2
PRE PROCESSING DAN ANALISIS



Next Page

PRE PROCESSING

```python
# Mengecek data duplikat
def check_duplicates(data, table_name):
    duplicate_count =data.duplicated().sum()
    print(f"\n--- Pengecekan Duplikat pada Tabel {table_name} ---")
    if duplicate_count > 0:
        print(f"Tabel {table_name} memiliki {duplicate_count} data duplikat.")
    else:
        print(f"Tabel {table_name} tidak memiliki data duplikat.")
    return duplicate_count

print ('Hasil Pengecekan Duplikat')

check_duplicates(data, "data")
```

```
Hasil Pengecekan Duplikat

--- Pengecekan Duplikat pada Tabel data ---
Tabel data memiliki 1 data duplikat.
1
```

```python
print("\nData Setelah Menghapus Duplikat:")
print(data)
```

```
Data Setelah Menghapus Duplikat:
      age     sex     bmi  children smoker     region      charges
0      19  female  27.900         0    yes  southwest  16884.92400
1      18    male  33.770         1     no  southeast   1725.55230
2      28    male  33.000         3     no  southeast   4449.46200
3      33    male  22.705         0     no  northwest  21984.47061
4      32    male  28.880         0     no  northwest   3866.85520
...   ...     ...     ...       ...    ...        ...         ...
1333   50    male  30.970         3     no  northwest  10600.54830
1334   18  female  31.920         0     no  northeast   2205.98080
1335   18  female  36.850         0     no  southeast   1629.83350
1336   21  female  25.800         0     no  southwest   2007.94500
1337   61  female  29.070         0    yes  northwest  29141.36030
```

```python
# Pengecekan terhadap NaN atau Missing Value
missing_values = data.isnull().sum()
descriptive_stats = data.describe()
data.types = data.dtypes
missing_values, descriptive_stats, data.types
```

```
<ipython-input-10-5f95eafcd4b4>:4: UserWarning: Pandas doesn't allow columns to be created
  data.types = data.dtypes
(age         0
 sex         0
 bmi         0
 children    0
 smoker      0
 region      0
 charges     0
 dtype: int64,
                age          bmi     children       charges
 count  1337.000000  1337.000000  1337.000000   1337.000000
 mean     39.222139    30.663452     1.095737  13279.121487
 std      14.044333     6.100468     1.205571  12110.359656
 min      18.000000    15.960000     0.000000   1121.873900
 25%      27.000000    26.290000     0.000000   4746.344000
 50%      39.000000    30.400000     1.000000   9386.161300
 75%      51.000000    34.700000     2.000000  16657.717450
 max      64.000000    53.130000     5.000000  63770.428010,
 age           int64
 sex          object
 bmi         float64
 children      int64
 smoker       object
 region       object
 charges     float64
 dtype: object)
```

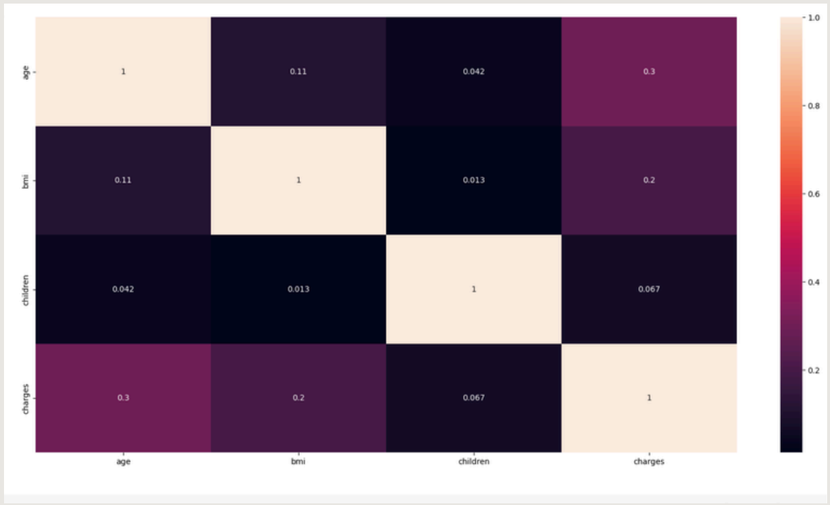
01

HASIL INTERPRETASI

Terdapatnya 1 Data Terduplikasi Dan
Tidak Terdapatnya *missing values* pada dataset

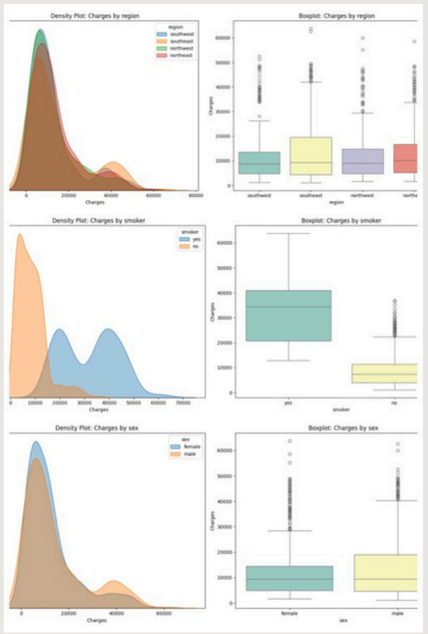
ANALISIS DESKRIPTIF

Matriks Korelasi

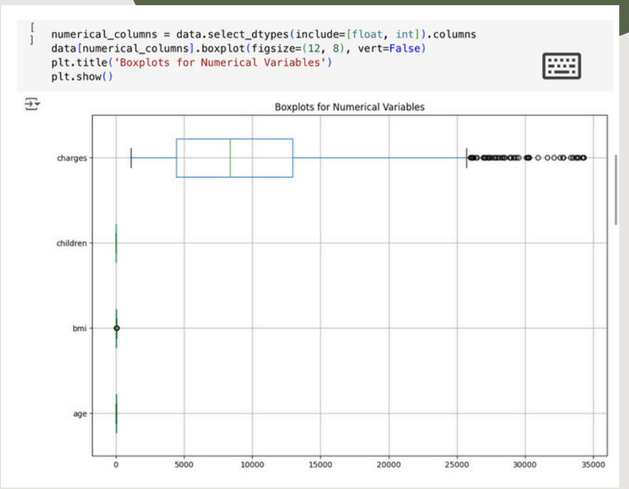


Dari heatmap dapat disimpulkan bahwa hubungan antar variabel dalam dataset menunjukkan korelasi yang lemah hingga sedang. Variabel age memiliki korelasi positif sedang dengan charges (0.3), yang menunjukkan bahwa semakin bertambah usia seseorang, biaya cenderung meningkat. Variabel BMI memiliki korelasi lemah dengan charges (0.2), mengindikasikan bahwa peningkatan BMI sedikit memengaruhi biaya, tetapi dampaknya kecil. Korelasi antara children dengan charges juga sangat lemah (0.067), menunjukkan bahwa jumlah anak hampir tidak memengaruhi biaya.

Selain itu, hubungan antar variabel independen seperti age, bmi, dan children semuanya sangat lemah (di bawah 0.15), menandakan bahwa tidak ada interaksi kuat di antara mereka.



Outliers



Outliers yang signifikan terlihat pada variabel charges, dengan nilai sangat tinggi di atas 50.000, kemungkinan besar berasal dari individu dengan status perokok (smoker=yes), BMI tinggi, atau kondisi kesehatan tertentu.

Untuk variabel numerik lainnya, seperti age, tidak terlihat outliers mencolok, sedangkan pada BMI terdapat outliers pada nilai ekstrem (>50), mencerminkan kasus obesitas parah. Pada variabel children, sebagian besar data berada di rentang 0-5 anak, tetapi ada beberapa kasus dengan jumlah anak lebih dari 5.

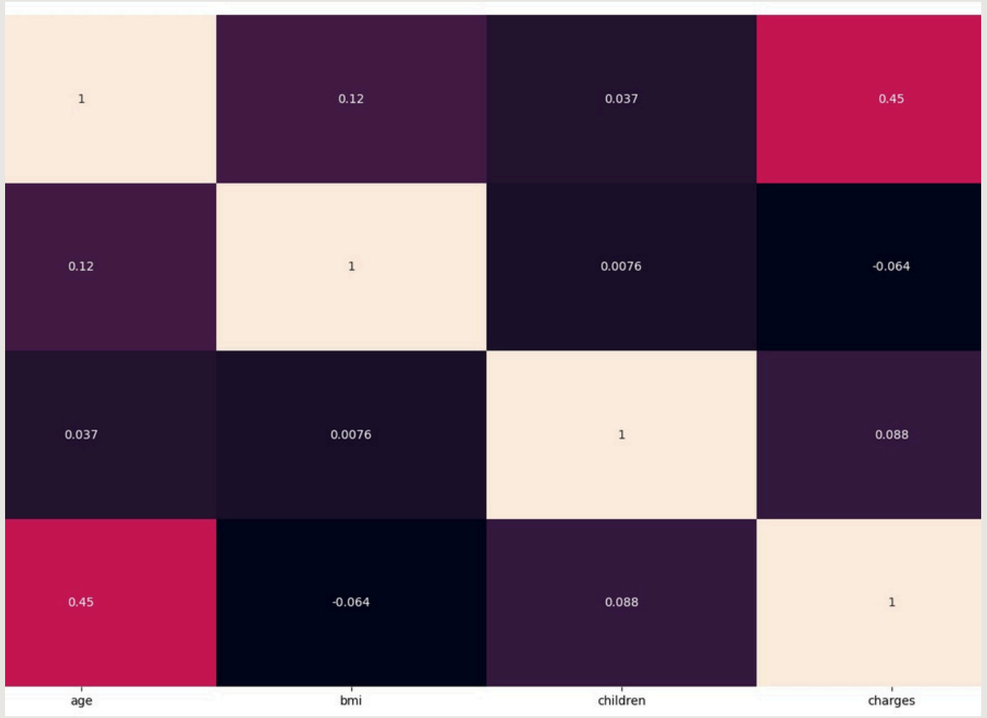
status perokok (smoker=yes) sangat berkontribusi terhadap nilai outliers pada charges. Interaksi antara usia, BMI tinggi, dan status perokok memainkan peran penting dalam menciptakan nilai-nilai ekstrem pada biaya kesehatan, sehingga outliers ini tidak perlu dihapus karena relevan dengan fenomena yang sedang dianalisis, meskipun penanganan seperti transformasi data diperlukan untuk analisis yang lebih akurat.

UNIVERSITAS INDONESIA

HYPOTHESIS

Biaya kesehatan dipengaruhi oleh berbagai faktor. Usia memiliki hubungan positif dengan biaya kesehatan karena risiko kesehatan meningkat seiring bertambahnya usia. BMI juga berpengaruh, di mana individu dengan BMI tinggi, terutama yang obesitas, cenderung memiliki biaya kesehatan lebih tinggi. Jumlah anak turut meningkatkan biaya, karena premi asuransi untuk keluarga menjadi lebih besar. Status perokok secara signifikan memengaruhi biaya kesehatan, dengan perokok memiliki pengeluaran lebih tinggi dibandingkan non-perokok. Jenis kelamin juga berperan, karena perbedaan risiko kesehatan spesifik antar gender. Selain itu, wilayah tempat tinggal memengaruhi biaya kesehatan, terutama karena kebijakan lokal yang dapat menaikkan biaya.

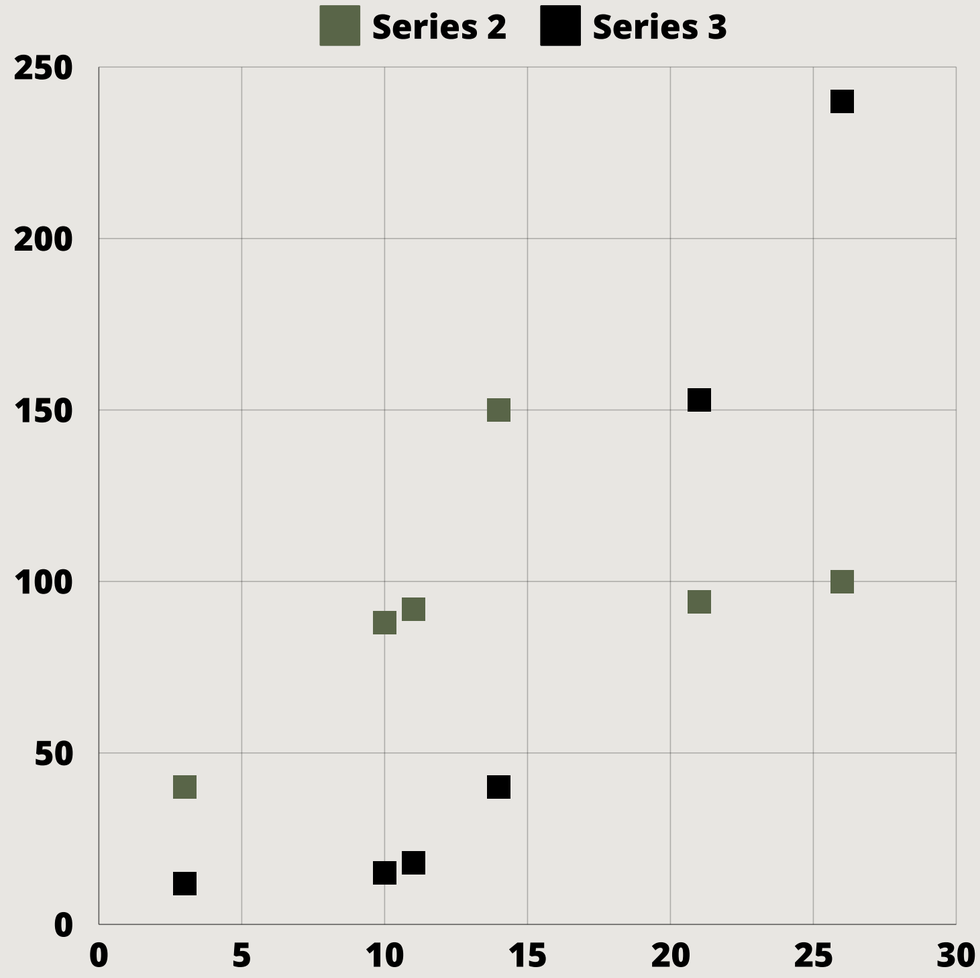
Terdapat beberapa interaksi antar variabel yang memengaruhi biaya kesehatan. Misalnya, individu yang lebih tua dan perokok cenderung memiliki biaya lebih besar dibandingkan individu muda yang bukan perokok. Individu dengan BMI tinggi yang juga merokok cenderung mengeluarkan biaya lebih tinggi dibandingkan yang tidak. Wilayah tempat tinggal juga berinteraksi dengan BMI, di mana individu obesitas di wilayah tertentu mungkin menghadapi biaya kesehatan yang lebih tinggi.



Next Page

UNIVERSITAS INDONESIA

BAGIAN 3
PEMODELAN



Next Page

MODEL 1

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_{smoker} + \beta_4 X_2 X_{smoker} + \epsilon$$

PENJELASAN NOTASI

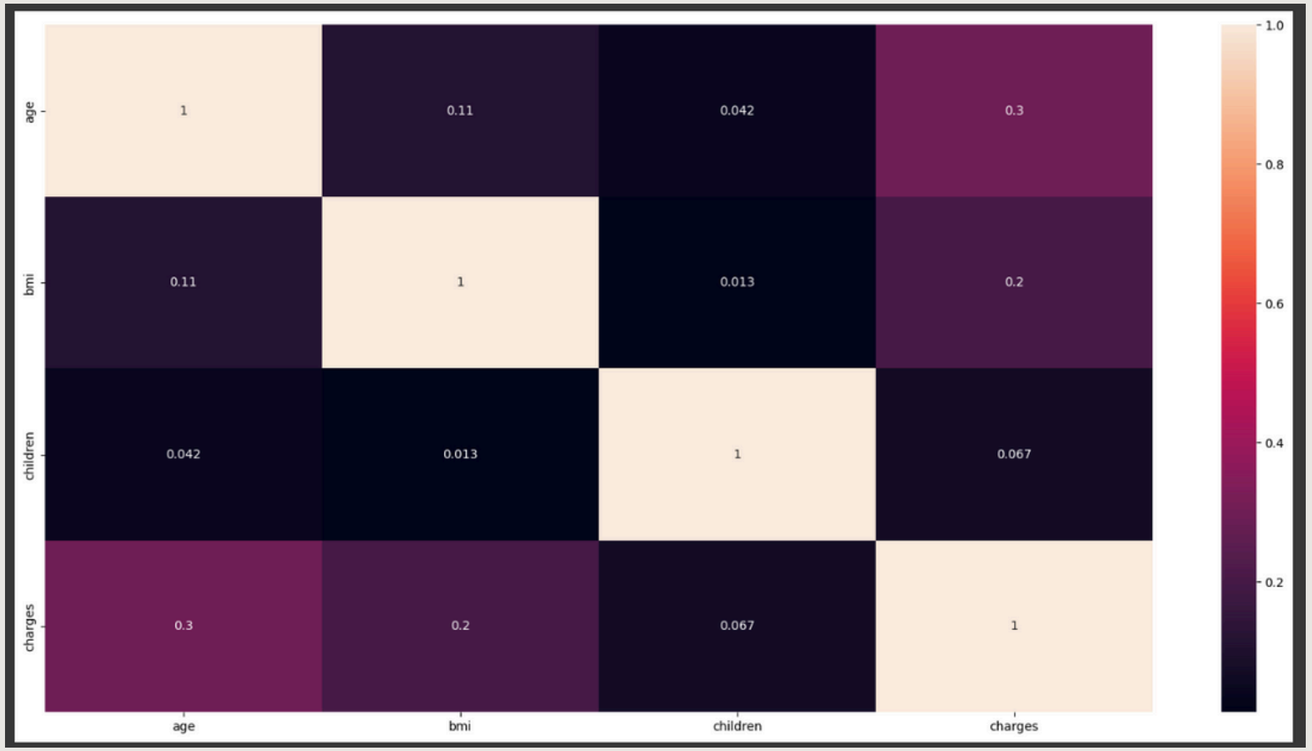
- Y adalah variabel respons (Charges)
- X_1 adalah variabel prediktor yang merepresentasikan umur / age (dalam tahun)
- X_2 adalah variabel prediktor yang merepresentasikan BMI (Body Mass Index)
- X_{smoker} adalah variabel prediktor dummy yang merepresentasikan apakah seorang adalah perokok (smoker) / tidak (variabel “smoker” adalah data kategorik dengan 2 level)
- $(\beta_2 + \beta_4 X_{smoker})$ menyatakan kecenderungan/tren rata-rata perubahan Y (Charges) saat X2 (BMI) naik sebesar satu satuan/unit, dengan variabel prediktor lain konstan
- $(\beta_3 + \beta_4 X_2)$ menyatakan kecenderungan/tren rata-rata perubahan Y (Charges) saat X_{smoker} (smoker) bernilai “1” (jika base levelnya adalah “No”, maka 3 menyatakan tren rata-rata perubahan Y (Charges) saat seseorang adalah perokok dan jika base levelnya adalah “Yes”, maka 3 menyatakan tren rata-rata perubahan Y (Charges) saat seseorang bukan perokok)
- β_2 , β_3, dan β_4 kurang memiliki interpretasi sendiri-sendiri

ASUMSI-ASUMSI

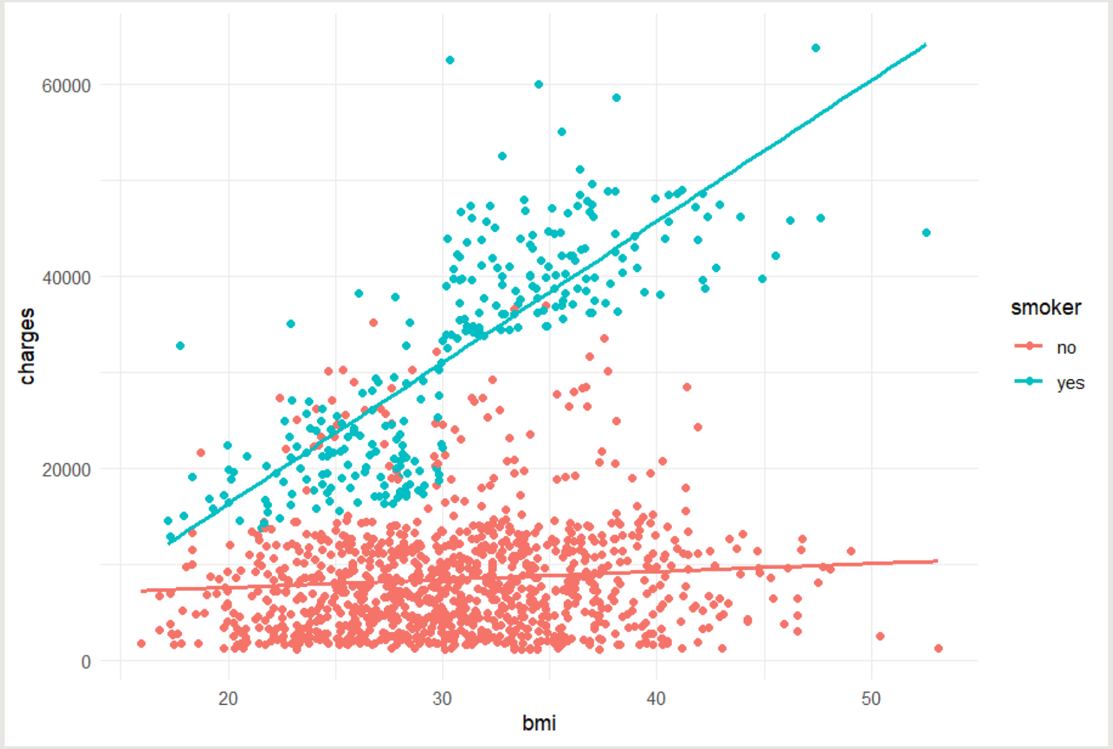
- Nilai dari variabel-variabel prediktor benar (tidak ada kesalahan pengukuran atau lainnya)
- Error memiliki mean 0 ($E(\varepsilon) = 0$)
- Error memiliki variansi konstan / homoskedastisitas ($Var(\varepsilon) = \sigma^2$)
- Error berdistribusi normal ($\sim N(0, \sigma^2)$)
- Error saling independen

ALASAN PENGAJUAN MODEL 1

BERKORELASI KUAT



TERLIHAT MEMILIK IINTERAKSI



MODEL 2

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_{smoker} + \beta_5 X_{sex} + \beta_6 X_{reg-1} + \beta_7 X_{reg-2} + \beta_8 X_{reg-3} + \epsilon$$

PENJELASAN NOTASI

- Y adalah variabel respons (Charges)
- X_1 adalah variabel prediktor yang merepresentasikan umur / age (dalam tahun)
- X_2 adalah variabel prediktor yang merepresentasikan BMI (Body Mass Index)
- X_3 adalah variabel prediktor yang merepresentasikan jumlah anak (Children)
- X_{smoker} adalah variabel prediktor dummy yang merepresentasikan apakah seorang adalah perokok (smoker) / tidak (variabel “smoker” adalah data kategorik dengan 2 level)
- X_{sex} adalah variabel prediktor dummy yang merepresentasikan jenis kelamin
- X_{rey-i}, i=1,2,3 adalah variabel prediktor dummy yang merepresentasikan daerah/region
- β_0 menyatakan rata-rata/ekspektasi nilai Y (Charges) saat umur, BMI, X_{smoker}, dan X bernilai 0
- β_1 menyatakan kecenderungan/tren rata-rata perubahan Y (Charges) saat X_1 (Age/umur) naik sebesar satu satuan/unit, dengan variabel prediktor lain konstan
- β_2 menyatakan kecenderungan/tren rata-rata perubahan Y (Charges) saat X_2 (BMI) naik sebesar satu satuan/unit, dengan variabel prediktor lain konstan
- β_3 menyatakan kecenderungan/tren rata-rata perubahan Y (Charges) saat X3 (Children / jumlah anak) naik sebesar satu satuan/unit, dengan variabel prediktor lain konstan
- β_4 menyatakan kecenderungan/tren rata-rata perubahan Y (Charges) saat X_{smoker} bernilai “1”
- β_5 menyatakan kecenderungan/tren rata-rata perubahan Y (Charges) saat X_{sex} bernilai “1”
- β_6, β_7, β_8 menyatakan kecenderungan/tren rata-rata perubahan Y (Charges) saat X_{rey-i} untuk i=1,2,3 secara berurut, bernilai “1"

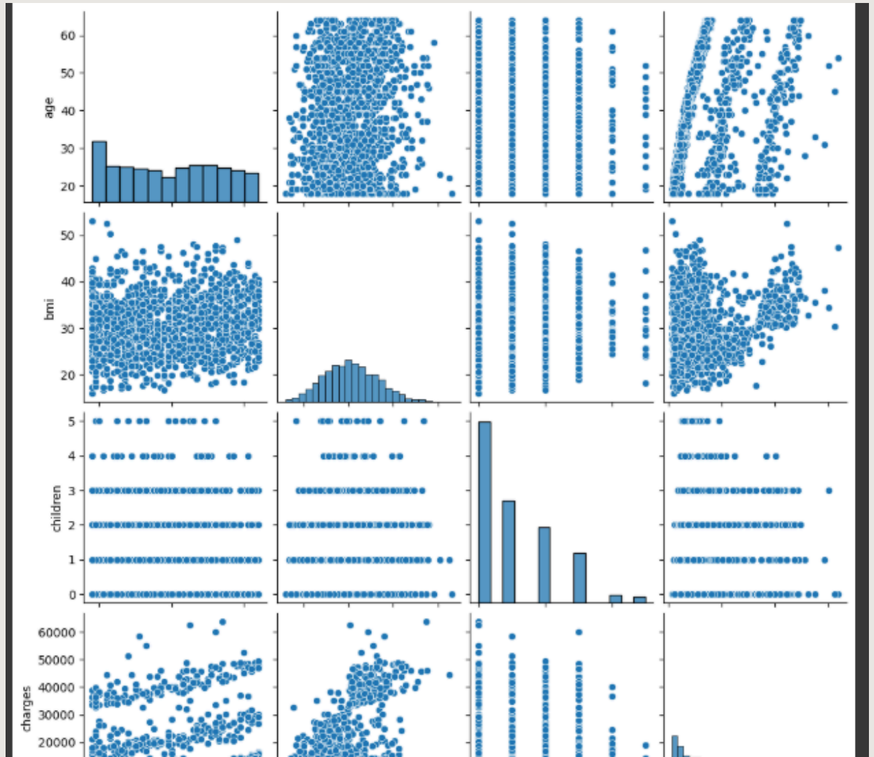
ASUMSI-ASUMSI

- Nilai dari variabel-variabel prediktor benar (tidak ada kesalahan pengukuran atau lainnya)
- Error memiliki mean 0 ($E(\varepsilon) = 0$)
- Error memiliki variansi konstan / homoskedastisitas ($Var(\varepsilon) = \sigma^2$)
- Error berdistribusi normal ($\sim N(0, \sigma^2)$)
- Error saling independen

ALASAN PENGAJUAN MODEL 2

TIDAK TERLIHAT INTERAKSI ORDE TINGGI

MENGUJI SIGNIFIKANSI PENAMBAHAN/ PENGURANGAN VARIABEL



RSS_1 = Residual Sum of Squares of fitted model 1

RSS_2 = Residual Sum of Squares of fitted model 2

$$F\ statistic = \frac{(\frac{RSS_1 - RSS_2}{k_2 - k_1})}{(\frac{RSS_2}{n - k_2})}$$

MODEL 3

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_{smoker} + \beta_5 X_{sex} + \beta_6 X_{reg-1} + \beta_7 X_{reg-2} + \beta_8 X_{reg-3} + \beta_9 X_2 X_{smoker} \epsilon$$

PENJELASAN NOTASI

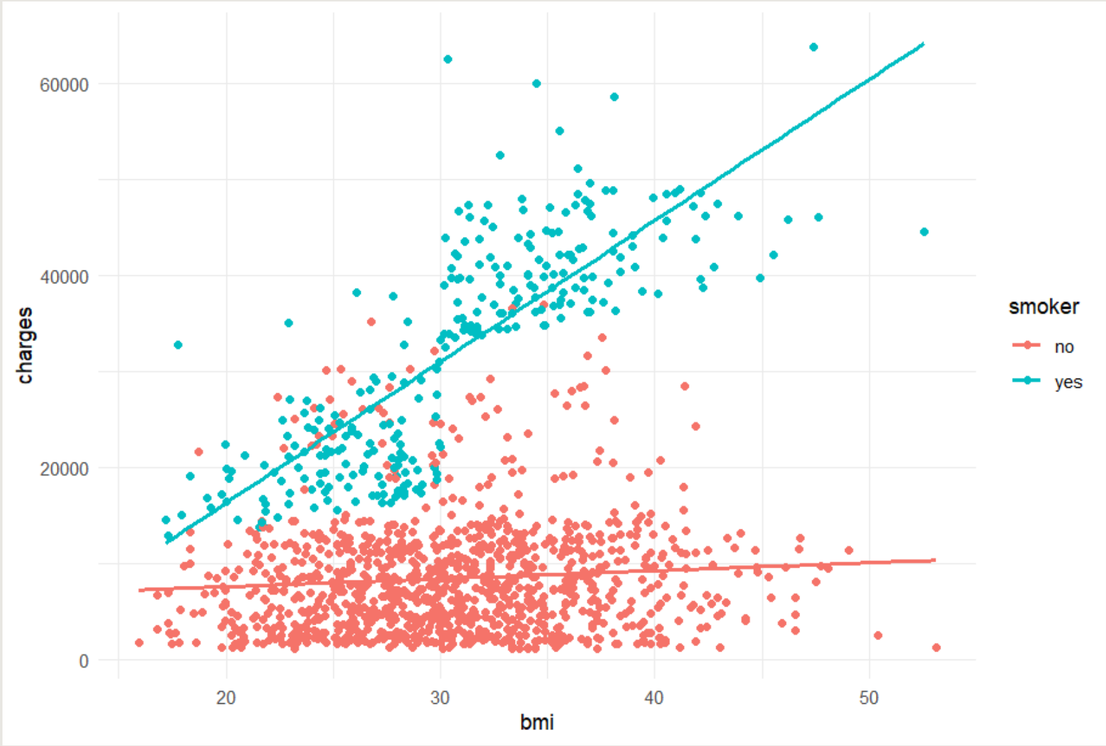
- Y adalah variabel respons (Charges)
- X_1, X_2, X_3, X_{smoker}, X_{sex}, dan X_{rey-i}, i=1,2,3 bermakna sama seperti di notasi Model 2
- β_0 menyatakan rata-rata/ekspektasi nilai Y (Charges) saat umur, BMI, X_{smoker}, dan X bernilai 0
- β_1 menyatakan kecenderungan/tren rata-rata perubahan Y (Charges) saat X_1 (Age/umur) naik sebesar satu satuan/unit, dengan variabel prediktor lain konstan
- (β_2+$\beta_9 X_{smoker}$) menyatakan kecenderungan/tren rata-rata perubahan Y (Charges) saat X_2 (BMI) naik sebesar satu satuan/unit, dengan variabel prediktor lain konstan
- β_3 menyatakan kecenderungan/tren rata-rata perubahan Y (Charges) saat X3 (Children / jumlah anak) naik sebesar satu satuan/unit, dengan variabel prediktor lain konstan
- (β_4+$\beta_9 X_2$) menyatakan kecenderungan/tren rata-rata perubahan Y (Charges) saat X_{smoker} bernilai “1”
- β_5 menyatakan kecenderungan/tren rata-rata perubahan Y (Charges) saat X_{sex} bernilai “1”
- β_6, β_7, β_8 menyatakan kecenderungan/tren rata-rata perubahan Y (Charges) saat X_{rey-i} untuk i=1,2,3 secara berurut, bernilai “1"
- β_2 , β_4, dan β_9 kurang memiliki interpretasi sendiri-sendiri

ASUMSI-ASUMSI

- Nilai dari variabel-variabel prediktor benar (tidak ada kesalahan pengukuran atau lainnya)
- Error memiliki mean 0 ($E(\varepsilon) = 0$)
- Error memiliki variansi konstan / homoskedastisitas ($Var(\varepsilon) = \sigma^2$)
- Error berdistribusi normal ($\sim N(0, \sigma^2)$)
- Error saling independen

ALASAN PENGAJUAN MODEL 3

INTERAKSI TERLIHAT SIGNIFIKAN



MENGUJI SIGNIFIKANSI PENAMBAHAN/ PENGURANGAN VARIABEL

RSS_1 = Residual Sum of Squares of fitted model 1

RSS_2 = Residual Sum of Squares of fitted model 2

$$F\ statistic = \frac{(\frac{RSS_1 - RSS_2}{k_2 - k_1})}{(\frac{RSS_2}{n - k_2})}$$

MODEL 4

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_{smoker} + \beta_5 X_{reg-1} + \beta_6 X_{reg-2} + \beta_7 X_{reg-3} + \beta_8 X_2 X_{smoker} \epsilon$$

PENJELASAN NOTASI

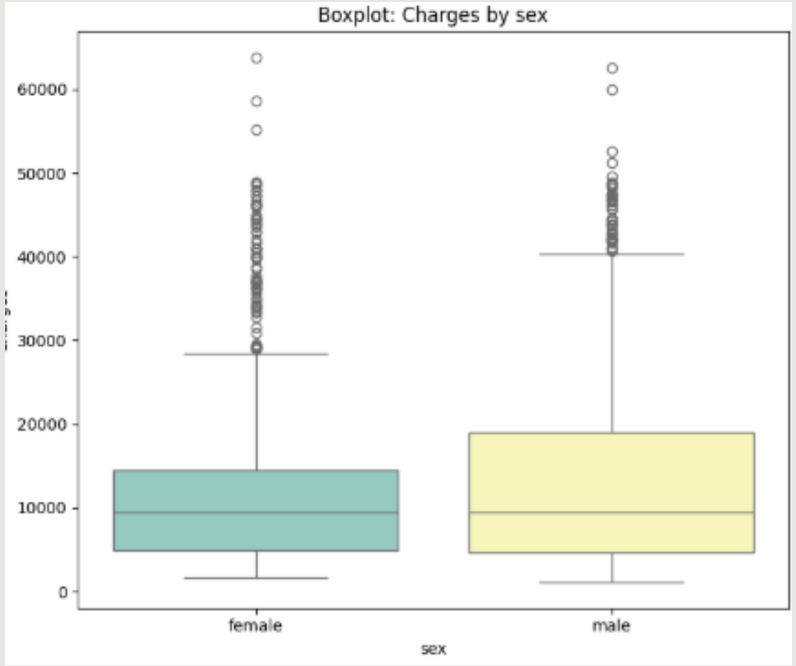
- Y adalah variabel respons (Charges)
- X_1, X_2, X_3, X_{smoker}, dan X_{rey-i}, i=1,2,3 bermakna sama seperti di notasi Model 2
- β_0 menyatakan rata-rata/ekspektasi nilai Y (Charges) saat umur, BMI, X_{smoker}, dan X bernilai 0
- β_1 menyatakan kecenderungan/tren rata-rata perubahan Y (Charges) saat X_1 (Age/umur) naik sebesar satu satuan/unit, dengan variabel prediktor lain konstan
- (β_2+$\beta_8 X_{smoker}$) menyatakan kecenderungan/tren rata-rata perubahan Y (Charges) saat X_2 (BMI) naik sebesar satu satuan/unit, dengan variabel prediktor lain konstan
- β_3 menyatakan kecenderungan/tren rata-rata perubahan Y (Charges) saat X3 (Children / jumlah anak) naik sebesar satu satuan/unit, dengan variabel prediktor lain konstan
- (β_4+$\beta_8 X_2$) menyatakan kecenderungan/tren rata-rata perubahan Y (Charges) saat X_{smoker} bernilai “1”
- β_5, β_6, β_7 menyatakan kecenderungan/tren rata-rata perubahan Y (Charges) saat X_{rey-i} untuk i=1,2,3 secara berurut, bernilai “1"
- β_2 , β_4, dan β_8 kurang memiliki interpretasi sendiri-sendiri

ASUMSI-ASUMSI

- Nilai dari variabel-variabel prediktor benar (tidak ada kesalahan pengukuran atau lainnya)
- Error memiliki mean 0 ($E(\varepsilon) = 0$)
- Error memiliki variansi konstan / homoskedastisitas ($Var(\varepsilon) = \sigma^2$)
- Error berdistribusi normal ($\sim N(0, \sigma^2)$)
- Error saling independen

ALASAN PENGAJUAN MODEL 4

PENGARUH JENIS KELAMIN TERLIHAT KURANG SIGNIFIKAN



MENGUJI SIGNIFIKANSI PENAMBAHAN/ PENGURANGAN VARIABEL

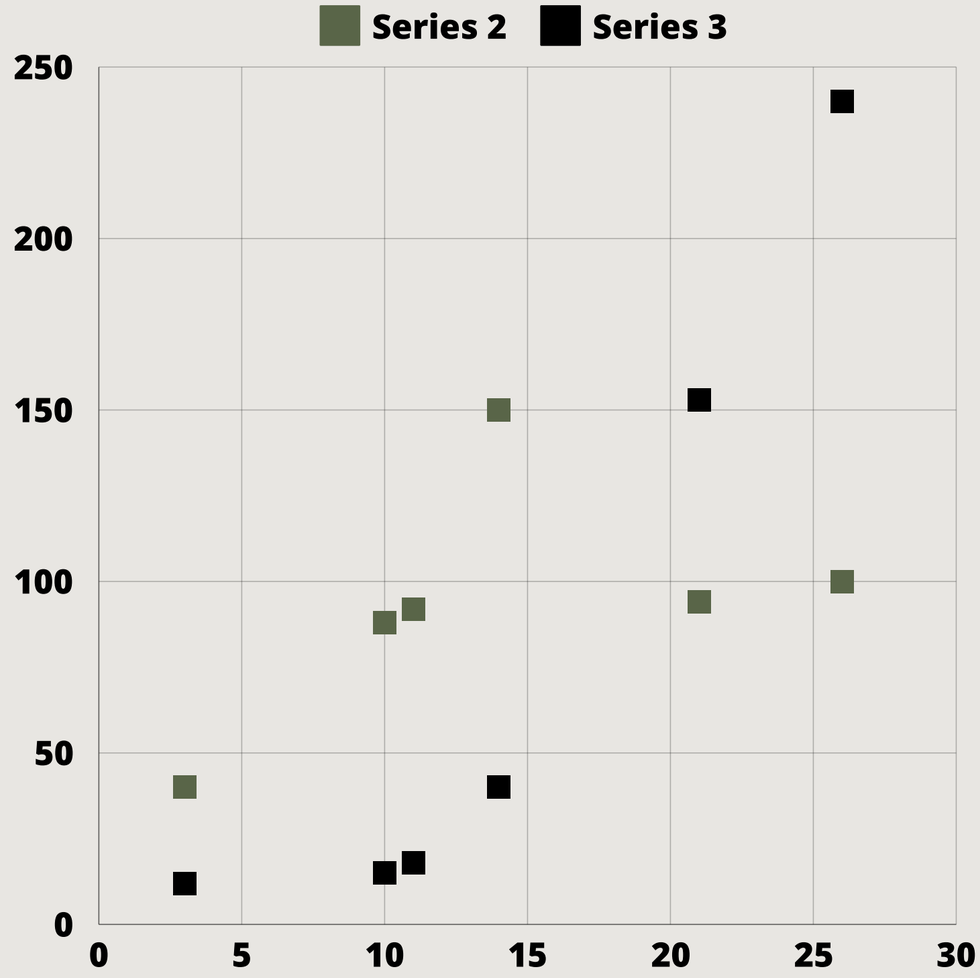
RSS_1 = Residual Sum of Squares of fitted model 1

RSS_2 = Residual Sum of Squares of fitted model 2

$$F\ statistic = \frac{(\frac{RSS_1 - RSS_2}{k_2 - k_1})}{(\frac{RSS_2}{n - k_2})}$$

UNIVERSITAS INDONESIA

BAGIAN 3
PEMODELAN

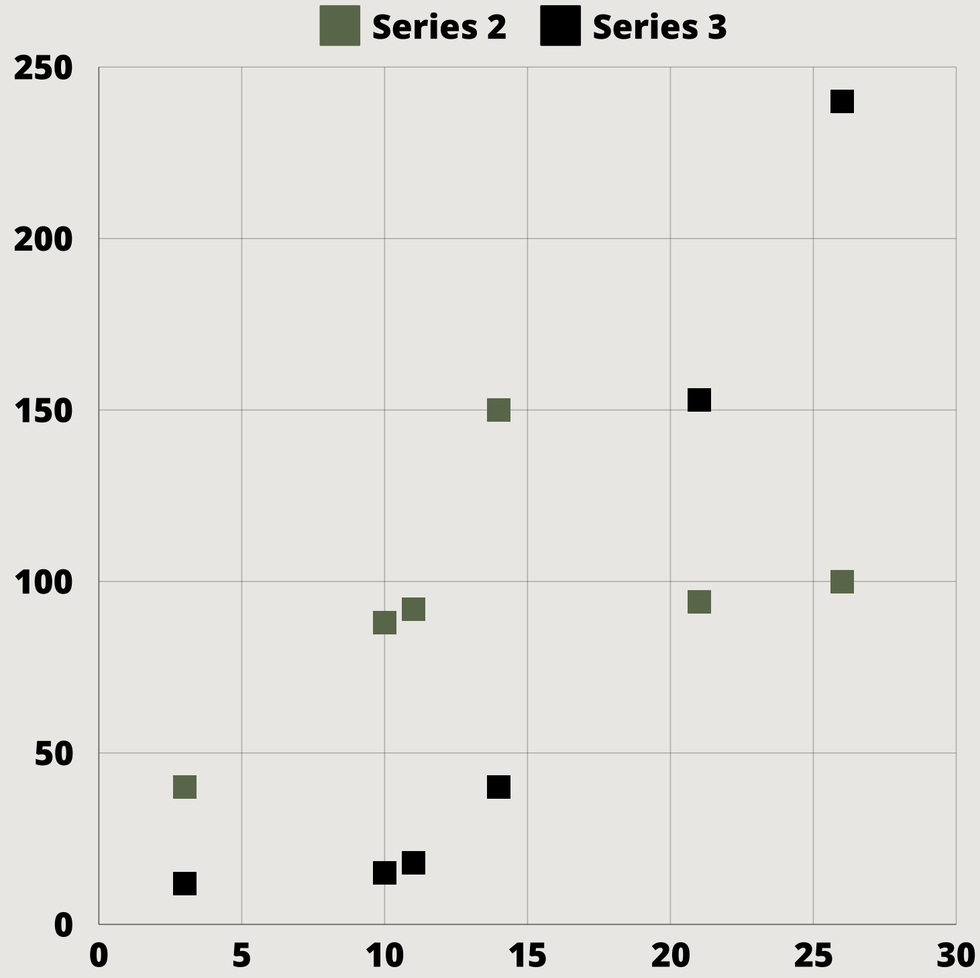


Next Page

UNIVERSITAS INDONESIA

BAGIAN 4

PENGOLAHAN DATA DAN ANALISIS HASIL



Next Page

4.1 PEMILIHAN MODEL BERDASARKAN KECUKUPAN MODEL

Setelah diajukan beberapa model pada bab 3, akan dilakukan validasi model terhadap masing-masing model yang telah diajukan. Kami melakukan pengolahan dan analisis data untuk mengevaluasi model yang diajukan. Analisis dimulai dengan melihat model summary untuk menginterpretasikan adjusted R-squared (Ra^2), yang menunjukkan seberapa besar variasi pada variabel respons dapat dijelaskan oleh variabel prediktor. Selanjutnya, uji-t dilakukan untuk menguji signifikansi masing-masing variabel independen, sementara uji-F digunakan untuk menilai signifikansi model secara keseluruhan.

MODEL PERTAMA

SUMMARY MODEL PERTAMA

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                charges   R-squared:                       0.836
Model:                            OLS   Adj. R-squared:                  0.836
Method:                 Least Squares   F-statistic:                     1702.
Date:                Sat, 21 Dec 2024   Prob (F-statistic):               0.00
Time:                        02:13:22   Log-Likelihood:                -13267.
No. Observations:                1338   AIC:                         2.654e+04
Df Residuals:                    1333   BIC:                         2.657e+04
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        -2290.0080    831.999     -2.752      0.006   -3922.179    -657.837
age            266.7582      9.617     27.739      0.000     247.893     285.624
bmi              7.1093     25.058      0.284      0.777     -42.049      56.267
Smoker_Dummy -2.009e+04   1666.827    -12.055      0.000   -2.34e+04   -1.68e+04
BMI_Smoker    1430.9204     53.217     26.888      0.000    1326.522    1535.319
==============================================================================
Omnibus:                      708.403   Durbin-Watson:                   2.062
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4237.479
Skew:                           2.485   Prob(JB):                         0.00
Kurtosis:                      10.162   Cond. No.                         660.
==============================================================================
```

```
Breusch-Pagan Test: p-value = 0.23738677929465393
        Features        VIF
0          const  38.459302
1            age   1.013536
2            bmi   1.296388
3   Smoker_Dummy  25.137153
4     BMI_Smoker  25.444810
R-squared: 0.8362767007435237
Adjusted R-squared: 0.835785408022574
```

Koefisien Determinasi
Didapatkan nilai *adjusted R-squared* dari model 1 yaitu 83,6%

Uji F (dengan Taraf Signifikansi $\alpha = 0.05$)

Hipotesis

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

H_1 : setidaknya terdapat satu parameter $\beta_i \neq 0$ untuk i = 1,2,3,4

Dengan p-value > α maka model dikatakan tidak *statistically useful.*

Uji T
Hipotesis

Hipotesis

$H_0 : \beta_i = 0$, untuk i = 1,2,3,4

$H_1 : \beta_i \neq 0$, untuk i = 1,2,3,4

dengan p-value < untuk setiap variabel kecuali variabel *bmi* sehingga variabel *bmi* tidak memiliki pengaruh terhadap variabel respons.

MODEL KEDUA

SUMMARY MODEL KEDUA

```
Kolom tersedia dalam dataset: Index(['age', 'sex', 'bmi', 'children', 'smoker', 'region', 'charges'], dtype='object')
                            OLS Regression Results
==============================================================================
Dep. Variable:                charges   R-squared:                       0.120
Model:                            OLS   Adj. R-squared:                  0.118
Method:                 Least Squares   F-statistic:                     60.69
Date:                Sat, 21 Dec 2024   Prob (F-statistic):           8.80e-37
Time:                        02:15:35   Log-Likelihood:                -14392.
No. Observations:                1338   AIC:                         2.879e+04
Df Residuals:                    1334   BIC:                         2.881e+04
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -6916.2433   1757.480     -3.935      0.000   -1.04e+04   -3468.518
age          239.9945     22.289     10.767      0.000     196.269     283.720
bmi          332.0834     51.310      6.472      0.000     231.425     432.741
children     542.8647    258.241      2.102      0.036      36.261    1049.468
Smoker_Dummy        0          0        nan        nan           0           0
==============================================================================
Omnibus:                      325.395   Durbin-Watson:                   2.012
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              603.372
Skew:                           1.520   Prob(JB):                    9.54e-132
Kurtosis:                       4.255   Cond. No.                          inf
==============================================================================
```

```
Breusch-Pagan Test: p-value = 4.77619570219101e-28
       Features        VIF
0         const  31.954929
1           age   1.013816
2           bmi   1.012152
3      children   1.001874
4  Smoker_Dummy        NaN
R-squared: 0.1200981957624696
Adjusted R-squared: 0.11811940609776739
```

Koefisien Determinasi
Didapatkan nilai *adjusted R-squared* dari model 2 yaitu 11,8%

Uji F (dengan Taraf Signifikansi $\alpha = 0.05$)

Hipotesis

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$

H_1 : setidaknya terdapat satu parameter $\beta_i \neq 0$ untuk i = 1,2,3,...,8.

Dengan p-value < α maka model dikatakan *statistically useful.*

Uji T
Hipotesis

Hipotesis

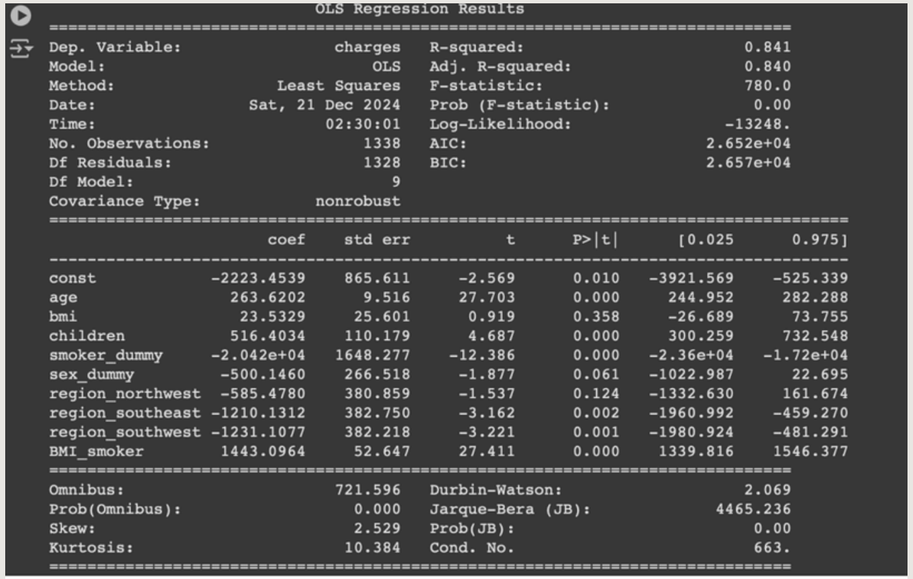
$H_0 : \beta_i = 0$, untuk i = 1,2,3,...,8.

$H_1 : \beta_i \neq 0$, untuk i = 1,2,3,...,8.

dengan p-value < untuk setiap variabel sehingga setiap variabel memiliki pengaruh terhadap variabel respons.

MODEL KETIGA

SUMMARY MODEL KETIGA



OLS Regression Results

Dep. Variable:	charges	R-squared:	0.841
Model:	OLS	Adj. R-squared:	0.840
Method:	Least Squares	F-statistic:	780.0
Date:	Sat, 21 Dec 2024	Prob (F-statistic):	0.00
Time:	02:30:01	Log-Likelihood:	-13248.
No. Observations:	1338	AIC:	2.652e+04
Df Residuals:	1328	BIC:	2.657e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2223.4539	865.611	-2.569	0.010	-3921.569	-525.339
age	263.6202	9.516	27.703	0.000	244.952	282.288
bmi	23.5329	25.601	0.919	0.358	-26.689	73.755
children	516.4034	110.179	4.687	0.000	300.259	732.548
smoker_dummy	-2.042e+04	1648.277	-12.386	0.000	-2.36e+04	-1.72e+04
sex_dummy	-500.1460	266.518	-1.877	0.061	-1022.987	22.695
region_northwest	-585.4780	380.859	-1.537	0.124	-1332.630	161.674
region_southeast	-1210.1312	382.750	-3.162	0.002	-1960.992	-459.270
region_southwest	-1231.1077	382.218	-3.221	0.001	-1980.924	-481.291
BMI_smoker	1443.0964	52.647	27.411	0.000	1339.816	1546.377

Omnibus:	721.596	Durbin-Watson:	2.069
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4465.236
Skew:	2.529	Prob(JB):	0.00
Kurtosis:	10.384	Cond. No.	663.

Breusch-Pagan Test: p-value = 0.3426653876376516

	Features	VIF
0	const	42.683356
1	age	1.017506
2	bmi	1.387358
3	children	1.004195
4	smoker_dummy	25.203045
5	sex_dummy	1.011478
6	region_northwest	1.519577
7	region_southeast	1.652690
8	region_southwest	1.530436
9	BMI_smoker	25.533081

R-squared: 0.8409179940508522

Adjusted R-squared: 0.8398398780466788

Koefisien Determinasi

Didapatkan nilai *adjusted R-squared* dari model 3 yaitu 84%

Uji F (dengan Taraf Signifikansi $\alpha = 0.05$)

Hipotesis

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$

H_1 : setidaknya terdapat satu parameter $\beta_i \neq 0$ untuk i = 1,2,3,...,9.

Dengan p-value > α maka model dikatakan tidak *statistically useful.*

Uji T

Hipotesis

Hipotesis

$H_0 : \beta_i = 0$, untuk i = 1,2,3,...9.

$H_1 : \beta_i \neq 0$, untuk i = 1,2,3,...,9.

dengan p-value < untuk setiap variabel kecuali variabel *children*, *sex_dummy*, dan *region_northwest* sehingga variabel tersebut tidak memiliki pengaruh terhadap variabel respons.

MODEL KEEMPAT

SUMMARY MODEL KEEMPAT

```
[ ]                            OLS Regression Results
==============================================================================
Dep. Variable:                charges   R-squared:                       0.840
Model:                            OLS   Adj. R-squared:                  0.840
Method:                 Least Squares   F-statistic:                     875.4
Date:                Sat, 21 Dec 2024   Prob (F-statistic):               0.00
Time:                        04:08:57   Log-Likelihood:                -13250.
No. Observations:                1338   AIC:                         2.652e+04
Df Residuals:                    1329   BIC:                         2.656e+04
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          -2453.5639    857.695     -2.861      0.004   -4136.147    -770.981
age              264.0422      9.522     27.729      0.000     245.362     282.723
bmi               22.6149     25.620      0.883      0.378     -27.645      72.875
children         512.7134    110.266      4.650      0.000     296.398     729.028
smoker_dummy  -2.031e+04   1648.861    -12.317      0.000   -2.35e+04   -1.71e+04
region_northwest  -581.7043    381.215     -1.526      0.127   -1329.554     166.145
region_southeast -1207.0113    383.109     -3.151      0.002   -1958.576    -455.446
region_southwest -1227.6015    382.576     -3.209      0.001   -1978.120    -477.083
BMI_smoker        1438.1084     52.630     27.325      0.000    1334.862    1541.355
==============================================================================
Omnibus:                      719.320   Durbin-Watson:                   2.072
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4430.079
Skew:                           2.521   Prob(JB):                         0.00
Kurtosis:                      10.352   Cond. No.                         662.
==============================================================================
```

```
Breusch-Pagan Test: p-value = 0.2588892344899111
            Features        VIF
0              const  41.826828
1                age   1.016938
2                bmi   1.386852
3           children   1.003875
4      smoker_dummy  25.173157
5  region_northwest   1.519534
6  region_southeast   1.652659
7  region_southwest   1.530399
8        BMI_smoker  25.467999
R-squared: 0.8404961372769608
Adjusted R-squared: 0.8395359936337823
```

Koefisien Determinasi

Didapatkan nilai *adjusted R-squared* dari model 4 yaitu 84%

Uji F (dengan Taraf Signifikansi $\alpha = 0.05$)

Hipotesis

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$

H_1 : setidaknya terdapat satu parameter $\beta_i \neq 0$ untuk i = 1,2,3,...,8.

Dengan p-value > α maka model dikatakan tidak *statistically useful.*

Uji T

Hipotesis

Hipotesis

$H_0 : \beta_i = 0$, untuk i = 1,2,3,...,8.

$H_1 : \beta_i \neq 0$, untuk i = 1,2,3,...,8.

dengan p-value < untuk setiap variabel kecuali variabel *smoker_dummy* dan *BMI_smoker* sehingga variabel tersebut tidak memiliki pengaruh terhadap variabel respons.

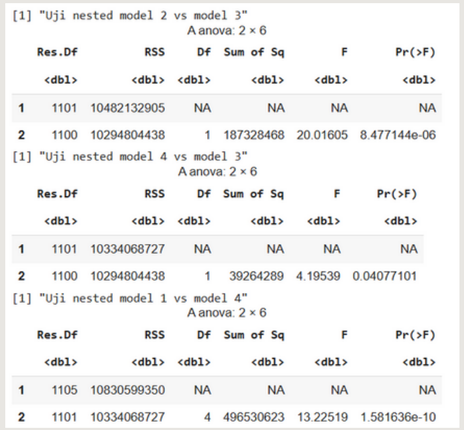
KESIMPULAN

Berdasarkan analisis uji kecukupan model yang telah dilakukan, kami memperoleh hasil bahwa model ketiga memiliki nilai adjusted R-squared yang paling tinggi, yaitu 0,840, namun model ini tidak memenuhi kriteria statistikal yang berguna karena H0 tidak dapat ditolak pada uji F. Hasil uji T juga menunjukkan bahwa sebagian besar variabel tidak dapat menjelaskan variabel respons secara signifikan, terutama variabel children, sex_dummy, dan region_northwest. Selain itu, terdapat masalah multikolinearitas pada variabel smoker_dummy dan BMI_smoker dengan nilai VIF yang lebih besar dari 10.

Model kedua, meskipun adjusted R-squared rendah (0,118), menunjukkan bahwa semua variabel yang dimasukkan memberikan pengaruh signifikan. Hasil uji F menunjukkan bahwa model ini secara keseluruhan dapat dijelaskan dengan baik oleh variabel prediktor. Untuk memastikan model terbaik, akan dilakukan uji perbandingan model (Nested Model) antara model. Oleh karena itu, akan dilanjutkan dengan analisis residual serta validasi model untuk memastikan kelayakan dan keakuratan model yang paling baik.

KONFIRMASI MODEL

UJI NESTED MODEL



[1] "Uji nested model 2 vs model 3"

Anova: 2 × 6

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1101	10482132905	NA	NA	NA	NA
2	1100	10294804438	1	187328468	20.01605	8.477144e-06

[1] "Uji nested model 4 vs model 3"

Anova: 2 × 6

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1101	10334068727	NA	NA	NA	NA
2	1100	10294804438	1	39264289	4.19539	0.04077101

[1] "Uji nested model 1 vs model 4"

Anova: 2 × 6

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1105	10830599350	NA	NA	NA	NA
2	1101	10334068727	4	496530623	13.22519	1.581636e-10

- Penambahan term interaksi berperan signifikan menjelaskan variansi dalam model. Model 3 > Model 2.
- Penambahan jenis kelamin kurang berperan signifikan menjelaskan variansi dalam model. Berdasarkan prinsip parsimoni, Model 4 > Model 3.
- Penambahan region dan children berperan signifikan menjelaskan variansi dalam model. Model 4 > Model 1

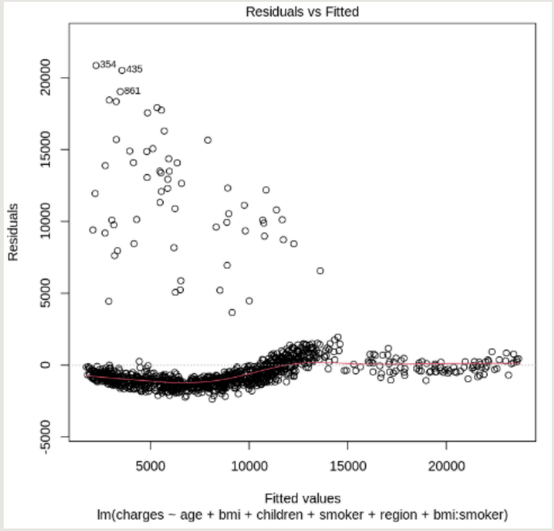
NILAI STATISTIK

	r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	0.685	0.684	3131.	601.	2.41e-275	4	-10507.	21026.	21056.
2	0.695	0.693	3086.	314.	9.22e-278	8	-10489.	20998.	21048.
3	0.701	0.698	3059.	286.	8.53e-281	9	-10479.	20980.	21035.
4	0.699	0.697	3064.	320.	3.73e-281	8	-10481.	20982.	21032.

- Model 3 adalah model dengan performa terbaik secara keseluruhan, karena memiliki Adjusted R-squared tertinggi, sigma terendah, serta AIC dan BIC yang paling kecil.
- Model 4 dapat dipertimbangkan jika ingin mengutamakan prinsip parsimony (model lebih sederhana) karena performanya hampir setara dengan Model 3.

HETEROSKEDASTISITAS

PENDETEKSIAN




```
        studentized Breusch-Pagan test

data:  model
BP = 30.675, df = 8, p-value = 0.0001605
```

Berdasarkan plot residual dan Uji Breusch-Pagan Test dengan *p-value* 0.0001605 < 0.05. Sehingga dapat disimpulkan terjadi **Heteroskedastisitas**.

PENANGANAN

```
#define weights to use
wt <- 1 / lm(abs(model$residuals) ~ model$fitted.values)$fitted.values^2

#perform weighted least squares regression
model <- lm(charges ~ age + bmi + children + smoker + region + bmi:smoker, data = data, weights=wt)
```

```
Call:
lm(formula = charges ~ age + bmi + children + smoker + region +
    bmi:smoker, data = data, weights = wt)

Weighted Residuals:
     Min       1Q   Median       3Q      Max
-27.0839  -0.7116  -0.3955   0.0322   9.9685

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     7238.26     104.75  69.099   <2e-16 ***
age             3576.74      63.51  56.316   <2e-16 ***
bmi              154.36      97.38   1.585   0.1132
children         504.39      11.52  43.791   <2e-16 ***
smokeryes      14561.68     112.47 129.475   <2e-16 ***
regionnorthwest    89.35      28.24   3.164   0.0016 **
regionsoutheast  -647.27      35.67 -18.145   <2e-16 ***
regionsouthwest  -652.70      29.65 -22.016   <2e-16 ***
bmi:smokeryes    2324.36     116.77  19.905   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.1 on 1101 degrees of freedom
Multiple R-squared:  0.9516,    Adjusted R-squared:  0.9512
F-statistic:  2705 on 8 and 1101 DF,  p-value: < 2.2e-16
```

```
        studentized Breusch-Pagan test

data:  model
BP = 1.6532e-05, df = 8, p-value = 1
```

Menggunakan *Weighted Least Square*, didapatkan R-Squared dari model meningkat menjadi 0.9516. Selain itu, hasil *p-value* Breusch Pagan Test 1> 0.05 sehingga residu sudah terdistribusi dengan variansi yang sama.

INSIGHT

- Pengaruh Usia terhadap Biaya Kesehatan: Semakin tua seseorang, semakin tinggi biaya kesehatan yang dikeluarkan. Hal ini mungkin disebabkan oleh peningkatan risiko penyakit seiring bertambahnya usia.
- BMI dan Status Perokok: Orang dengan BMI lebih tinggi (indikasi kelebihan berat badan/obesitas) cenderung memiliki biaya kesehatan yang lebih besar. Interaksi BMI dan status perokok menunjukkan bahwa efek BMI terhadap biaya kesehatan lebih besar untuk perokok dibanding non-perokok.
- Status Perokok: Biaya kesehatan perokok jauh lebih tinggi dibandingkan non-perokok, bahkan setelah memperhitungkan variabel lain.
- Jumlah Anak: Jumlah anak memiliki dampak positif pada biaya kesehatan, meskipun pengaruhnya relatif kecil dibanding variabel lain. Ini mencerminkan pengaruh tanggungan keluarga dalam menentukan biaya asuransi atau kesehatan.
- Wilayah Tempat Tinggal: Ada perbedaan biaya kesehatan yang konsisten antara wilayah (misalnya, karena variasi biaya layanan medis, kebiasaan hidup, atau akses kesehatan).
- Jenis Kelamin Tidak Signifikan: Variabel jenis kelamin tidak menunjukkan pengaruh signifikan terhadap biaya kesehatan.
- Heteroskedastisitas: Masalah heteroskedastisitas yang ditemukan menunjukkan bahwa variabilitas biaya kesehatan lebih besar pada kelompok tertentu (misalnya, kelompok usia lebih tua atau perokok). Ini memberikan indikasi adanya risiko kesehatan yang lebih tinggi pada subkelompok tertentu.
- Transformasi Data untuk Stabilitas Model: Transformasi log pada variabel charges menunjukkan bahwa biaya kesehatan memiliki distribusi yang tidak normal, dengan sebagian kecil populasi memiliki biaya yang jauh lebih besar.

KESIMPULAN

$$Y = X_1 + X_2 + X_3 + X_{smoker} + X_{reg-1} + X_{reg-2} + X_{reg-3} + X_2X_{smoker} + \varepsilon$$

Notasi:

- Y: Charges
- X1: Age
- X2: BMI
- X3: Jumlah anak (children)
- Xsmoker: Smoker (variabel dummy dari data kategorik 2 level)
- Xreg-i , i=1,2,3: Region (variabel dummy dari data kategorik 4 level)

Pemilihan ini didasarkan oleh beberapa alasan yaitu:

- Adjusted R-squared tertinggi, menunjukkan model ini memiliki kemampuan menjelaskan variabilitas data yang lebih baik dibandingkan model lainnya.
- AIC dan BIC terendah, menunjukkan efisiensi model dengan jumlah parameter yang digunakan.
- Prinsip parsimoni, di mana model ini memberikan keseimbangan antara kompleksitas dan kualitas fit.

Model yang dihasilkan (Model 4) berhasil menjawab permasalahan yang diajukan dengan menghasilkan prediksi biaya kesehatan yang akurat dan penjelasan yang mendalam tentang faktor-faktor yang memengaruhi biaya tersebut. Dengan mempertimbangkan interaksi variabel seperti BMI dan status perokok, model ini memberikan wawasan penting untuk pengambilan keputusan terkait kesehatan dan kebijakan asuransi.