UNIVERSITAS TERBUKA

Tugas ke-3

Pengantar Sains Data

STDA4101

Nama	**:**	**MOH EVAN ALSADIK**
NIM	**:**	**053779188**
Kelas	**:**	**STDA4101.1**
UPBJJ-UT	**:**	**Bandung**

Dosen Pengampu Mata Kuliah:

Dimas Agung Prasetyo, S.Kom., M.S. | 01001664

Sains Data

Fakultas Teknologi dan Sains

2024

Soal & Pembahasan

Berdasarkan data ukuran kuantitatif perkembangan penyakit setelah satu tahun amatan (Y), usia dalam tahun (AGE), body mass index (BMI), average blood pressure (BP), dan total serum cholesterol (S1) terlampir, lakukan analisis statistik deskriptif dengan cara:

1. **Buatlah visualisasi data multivariat menggunakan Scatter Plot**
2. **Tentukan model persamaan regresi linear**
3. **Berikan ulasan Anda tentang model tersebut.**

Jawaban Tugas 3 STDA4101-24.1 (jawaban 1 dan 2)

December 10, 2024

Nama: Moh Evan Alsadik | Nim: 053779188 | STDA4101

[54]:
```
from scipy.stats import linregress
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import seaborn as sns
import statsmodels.api as sm
file_tugas = "Tugas 3 STDA4101-2024.1.xlsx"
tugas_df = pd.read_excel(file_tugas)
# tugas_df.describe()

age = tugas_df["AGE"]
body_mass_index = tugas_df["BMI"]
average_blood_pressure = tugas_df["BP"]
total_serum_cholesterol = tugas_df["S1"]
perkembangan_penyakit = tugas_df["Y"]
print("\n")
```

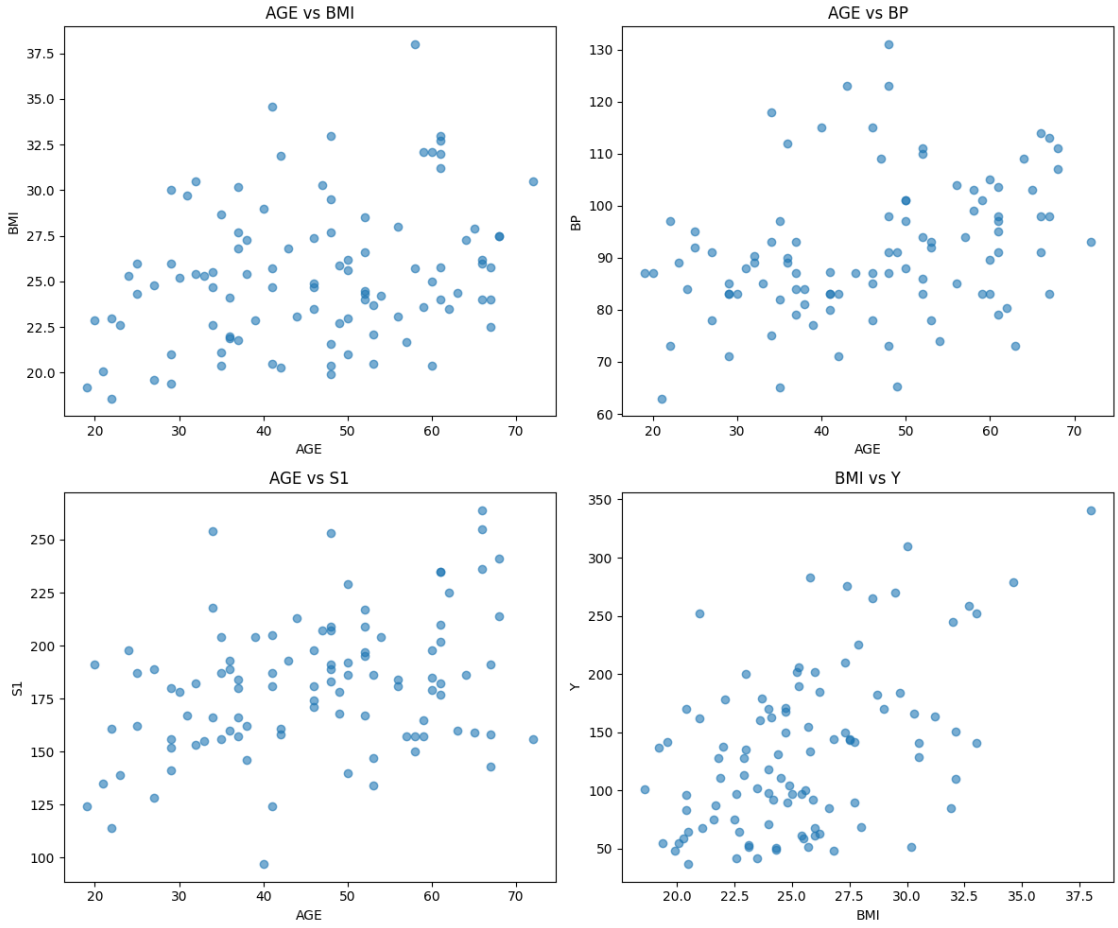
Berdasarkan data ukuran kuantitatif perkembangan penyakit setelah satu tahun amatan (Y), usia dalam tahun (AGE), body mass index (BMI), average blood pressure (BP), dan total serum cholesterol (S1) terlampir, lakukan analisis statistik deskriptif dengan cara:

0.0.1 Buatlah visualisasi data multivariat menggunakan Scatter Plot.

Data multivariat merupakan data yang dikumpulkan dari dua atau lebih pengamatan dan diukur menggunakan beberapa karakteristik. Singkatnya, data ini melibatkan lebih dari satu variabel baik independent maupun dependent yang dilakuklan analisis secara bersamaan.

Scatterplot adalah salah satu grafik yang umum digunakan dalam menemukan suatu pola hubungan antara 2 variabel. Komponen pada grafik ini terdiri dari garis memanjang (x) sebagai variabel independent, garis melebar untuk variabel dependen, serta titik yang merepresentasikan nilai x dan y.

[52]:
```
# Grid 2x2
fig, axes = plt.subplots(2, 2, figsize=(12, 10))

# Plot AGE dengan BMI
axes[0, 0].scatter(age, body_mass_index, alpha=0.6)
axes[0, 0].set_title('AGE vs BMI')
axes[0, 0].set_xlabel('AGE')
axes[0, 0].set_ylabel('BMI')

# Plot AGE dengan BP
axes[0, 1].scatter(age, average_blood_pressure, alpha=0.6)
axes[0, 1].set_title('AGE vs BP')
axes[0, 1].set_xlabel('AGE')
axes[0, 1].set_ylabel('BP')

# Plot AGE dengan S1
axes[1, 0].scatter(age, total_serum_cholesterol, alpha=0.6)
axes[1, 0].set_title('AGE vs S1')
axes[1, 0].set_xlabel('AGE')
axes[1, 0].set_ylabel('S1')

# Plot BMI dengan Y
axes[1, 1].scatter(body_mass_index, perkembangan_penyakit, alpha=0.6)
axes[1, 1].set_title('BMI vs Y')
axes[1, 1].set_xlabel('BMI')
axes[1, 1].set_ylabel('Y')

plt.tight_layout()  # Atur tata letak supaya tidak tumpang tindih
plt.show()
print("\n")
```



0.0.2 Tentukan model persamaan regresi linear.

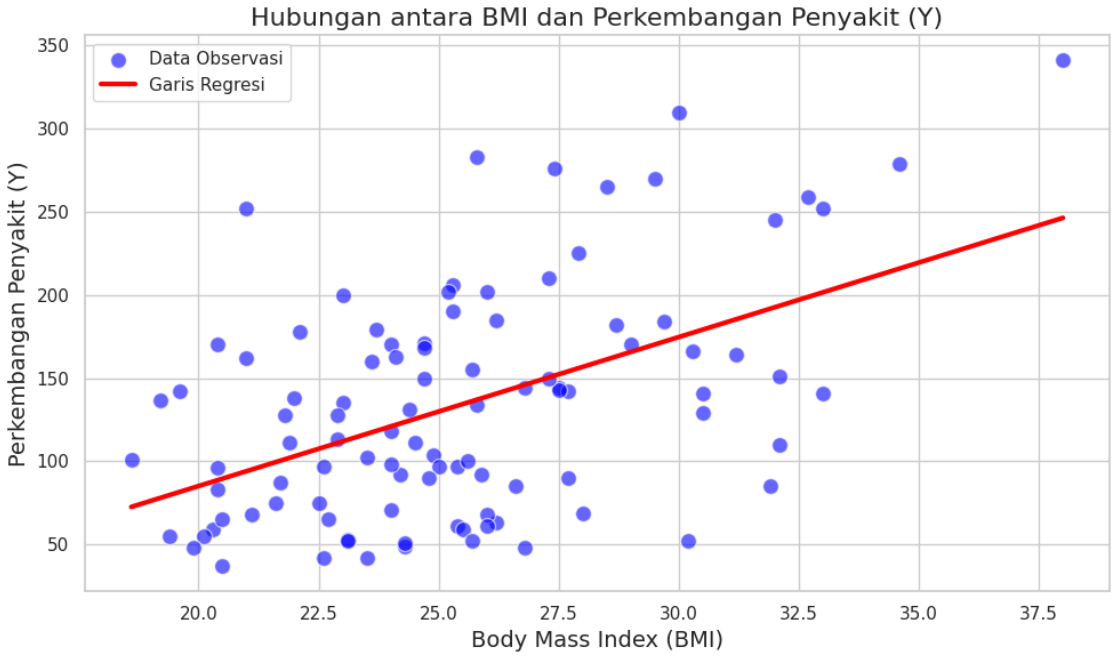
[53]:
```
# Variabel independen (AGE, BMI, BP, S1)
X = tugas_df[['AGE', 'BMI', 'BP', 'S1']]

# Menambahkan konstanta (intercept) ke variabel independen
X = sm.add_constant(X)  # Tambahkan kolom konstanta (1) agar ada intercept

# Variabel dependen (Y)
Y = tugas_df['Y']

# Model regresi
model = sm.OLS(Y, X).fit()

# Ringkasan hasil regresi
```

```
print(model.summary())
print("\n")

# Menampilkan hasil regresi
print("Persamaan Regresi Linear:")
print(f"Y = {intercept:.2f} + {slope:.2f} * X")
print(f"Koefisien Korelasi (r): {r_value:.2f}")  # Menunjukkan kekuatan dan arah hubungan
print(f"p-value: \t\t{p_value:.4f}")  # Menguji signifikansi hubungan antar variabel
print(f"Standard Error: \t{std_err:.2f}")  # Mengukur ketidakpastian dalam estimasi slope.
print("\n")
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.270
Model:                            OLS   Adj. R-squared:                  0.239
Method:                 Least Squares   F-statistic:                     8.768
Date:                Tue, 10 Dec 2024   Prob (F-statistic):           4.55e-06
Time:                        11:42:49   Log-Likelihood:                -549.34
No. Observations:                 100   AIC:                             1109.
Df Residuals:                      95   BIC:                             1122.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -119.8912     53.293     -2.250      0.027    -225.690     -14.092
AGE           -0.0748      0.493     -0.152      0.880      -1.054       0.905
BMI            8.3701      1.728      4.843      0.000       4.939      11.801
BP             0.8582      0.517      1.659      0.100      -0.169       1.885
S1            -0.1886      0.213     -0.886      0.378      -0.611       0.234
==============================================================================
Omnibus:                        2.415   Durbin-Watson:                   1.986
Prob(Omnibus):                  0.299   Jarque-Bera (JB):                2.400
Skew:                           0.329   Prob(JB):                        0.301
Kurtosis:                       2.620   Cond. No.                     1.87e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.87e+03. This might indicate that there are strong multicollinearity or other numerical problems.
```

```
Persamaan Regresi Linear:
Y = -94.10 + 8.96 * X
Koefisien Korelasi (r): 0.50
p-value:                0.0000
Standard Error:         1.58
```

[55]:
```python
# Data untuk BMI dan Y
bmi = tugas_df['BMI']
y = tugas_df['Y']

# Komponen persamaan regresi
intercept = -94.10
slope = 8.96

# Membuat garis regresi berdasarkan persamaan yang telah didapat
bmi_sorted = np.sort(bmi)
y_pred = intercept + slope * bmi_sorted

# Membuat plot
plt.figure(figsize=(10, 6))
sns.set(style="whitegrid")

# Scatter Plot
plt.scatter(bmi, y, color='blue', alpha=0.6, label="Data Observasi",␣
 ↪edgecolors="w", s=100)

# Garis Regresi
plt.plot(bmi_sorted, y_pred, color='red', linewidth=3, label="Garis Regresi")

# Label dan Grid
plt.title("Hubungan antara BMI dan Perkembangan Penyakit (Y)", fontsize=16)
plt.xlabel("Body Mass Index (BMI)", fontsize=14)
plt.ylabel("Perkembangan Penyakit (Y)", fontsize=14)
plt.legend()
plt.grid(True)
plt.tight_layout()

# Menampilkan plot
plt.show()
print("\n")
```


Hubungan antara BMI dan Perkembangan Penyakit (Y)
Data Observasi
Garis Regresi
Perkembangan Penyakit (Y)
Body Mass Index (BMI)
350
300
250
200
150
100
50
20.0
22.5
25.0
27.5
30.0
32.5
35.0
37.5

Soal & Pembahasan

0.0.3 Berikan ulasan Anda tentang model tersebut

Keterangan mengenai definisi, komponen, dan pengujian signifikansi (pada tugas sebelumnya)

Persamaan regresi linear:

$$Y = a + bX$$

ket.

- a = intersep (nilai Y ketika X = 0) => intercept
- b = koefisien regresi (kemiringan/gradien) => slope
- Y = variabel dependen (perkembangan penyakit)
- X = variabel independen (usia)

p-value:

nilai probabilitas yang digunakan pada pengujian hipotesis. Nilai ini digunakan untuk mengukur apakah hubungan tersebut signifikan secara statistik atau hanya terjadi karena kebetulan.

- p<0.05: Hubungan antara variabel X dan Y *signifikan secara statistik*. Artinya, peluang hubungan tersebut terjadi secara kebetulan adalah kurang dari 5%.
- p>0.05: Tidak cukup bukti untuk menyatakan hubungan signifikan. Artinya ada indikasi bahwa lebih dari 5% hubungan yang ada terjadi secara kebetulan.

stderr (standard error):

ukuran variabilitas dalam estimasi parameter, sebagai contoh nilai slope (b). Nilai stderr ini memberi gambaran seberapa luas kemungkinan estimasi parameter bisa berubah jika data yang digunakan berbeda (misalnya, dalam pengambilan sampel data ulang).

- Jika nilai stderr kecil, berarti estimasi parameter stabil dan dapat dipercaya (karena variasi parameternya kecil).
- Jika nilai stderr besar, berarti estimasi parameter kurang akurat, sehingga hasil regresi menjadi diragukan.

Didapatkan **persamaan regresinya** adalah:

$$Y = -94.10 + 8.96X$$

persamaan ini didapatkan dari penggabungan beberapa variabel independen (yaitu **BMI**, **AGE**, **BP**, dan **S1**) untuk memprediksi variabel dependen (yaitu **Y**) menggunakan multiple linear regression. Hasil persamaan regresi yang sebenarnya adalah:

$$Y = -94.10 + 8.96 * BMI + (-0.0748 * AGE) + 0.8582 * BP + (-0.1886 * S1)$$

$$Y = -94.10 + 8.96 * BMI - 0.0748 * AGE + 0.8582 * BP - 0.1886 * S1$$

$$Y = -94.10 + 8.96 * BMI$$

$$Y = -94.10 + 8.96 * X$$

Hanya saja, karena variabel **BMI** menjadi satu-satunya variabel yang memberikan pengaruh yang signifikan (dilihat dari nilai intercept nya), maka variabel sisanya (**AGE**, **BP**, dan **S1**) tidak ikut dimasukkan. Hal ini dikarenakan ketiganya tidak memberikan pengaruh signifikan meskipun tetap digunakan dalam persamaan.

Keterangan:

1. Intercept dan Koefisien (Koefisien Regresi)

- **Intercept (-94.10):** Nilai ini menunjukkan bahwa jika semua variabel independen (AGE, BMI, BP, dan S1) adalah 0, maka nilai perkembangan penyakit (Y) diperkirakan -94.10 (meskipun nilai ini tidak realistis, karena kita tidak mungkin memiliki nilai 0 untuk variabel seperti **BMI**).

- **Koefisien BMI (8.96):** Koefisien untuk **BMI** menunjukkan bahwa setiap kenaikan satu unit pada **BMI** akan meningkatkan perkembangan penyakit (Y) sebesar **8.96** satuan. Berarti semakin tinggi **BMI**, semakin tinggi perkembangan penyakitnya (menunjukkan pengaruh positif **BMI** terhadap **Y**, yang).

Selanjutnya untuk:

1. **Signifikansi Koefisien**
    a. **p-value untuk BMI = 0.0000**: Menunjukkan bahwa koefisien **BMI** sangat signifikan, karena p-value lebih kecil dari 0.05 (ada hubungan yang kuat dan signifikan antara **BMI** dan **Y)**.
    b. **p-value untuk AGE (0.880)**: Menunjukkan bahwa **AGE** tidak memiliki pengaruh signifikan terhadap **Y**.
    c. **p-value untuk BP (0.100)**: Meskipun p-value untuk **BP** mendekati 0.05, hasil ini masih menunjukkan bahwa **BP** tidak secara signifikan mempengaruhi **Y** (mungkin masih memiliki pengaruh, tetapi tidak cukup kuat).
    d. **p-value untuk S1 (0.378)**: Pengaruh **S1** juga tidak signifikan terhadap **Y**. Nilai p-value yang lebih besar dari 0.05 menunjukkan bahwa **S1** tidak memiliki hubungan yang cukup kuat dengan **Y** dalam analisis.

2. **R-squared dan Nilai Adjusted R-squared**
    a. **R-squared (0.270):** Nilai ini menunjukkan bahwa model ini hanya dapat menjelaskan sekitar **27%** variasi dalam perkembangan penyakit (**Y**). Artinya ada banyak faktor lain yang mempengaruhi **Y**, selain dari variabel yang digunakan dalam model.
    b. **Adjusted R-squared (0.239):** Nilai yang mengoreksi R-squared dengan mempertimbangkan jumlah variabel independen. Nilai yang lebih rendah dari R-squared menandakan bahwa model mungkin tidak terlalu baik dalam menjelaskan variasi dalam **Y**, dan variabel lain yang tidak dimasukkan dalam model bisa jadi berpengaruh.

3. **Standar Error**

   Standard Error (1.58): Nilai ini mengukur sejauh mana estimasi koefisien regresi dapat bervariasi. Nilai yang lebih kecil menunjukkan estimasi koefisien yang lebih akurat. Maka dari itu model regresi ini bisa dibilang cukup akurat.

4. **F-statistic dan p-value F-statistic**

   F-statistic (8.768): Ini menguji apakah model secara keseluruhan signifikan dalam memprediksi **Y**. Karena nilai **p-value untuk F-statistic (4.55e-06)** sangat kecil, berarti model regresi secara keseluruhan adalah signifikan, meski hanya variabel **BMI** yang memberi kontribusi signifikan terhadap model.

Kesimpulan:

Berdasarkan hasil analisis regresi linear berganda, diperoleh persamaan regresi sebagai berikut:

$$Y = -94.10 + 8.96 * X$$

Meskipun persamaan awal melibatkan variabel **AGE**, **BP**, dan **S1**, hanya variabel **BMI** saja yang memiliki pengaruh signifikan terhadap perkembangan penyakit (**Y**) karena p-value BMI (0.0000) lebih kecil dari 0.05 dan paling kecil dibandingkan variabel lain. Oleh karena itu, hanya **BMI** yang dipertahankan dalam persamaan akhir. Meskipun modelnya signifikan secara keseluruhan, tetap saja kemampuan model ini dalam menjelaskan variasi **Y** hanya sebesar **27%**, sehingga perlu mempertimbangkan faktor lain dalam model prediksi.

Sumber Referensi

Sutikno., & Ratnaningsih, D. J. (2025). Metode Statistika I. Modul 01 & 02. Tangerang, Banten. Universitas Terbuka.

Wijaya, T., & Budiman, S. (2016). Analisis multivariat untuk penelitian manajemen. Yogyakarta: Pohon Cahaya.

https://learn.nural.id/course/statistics/regresi-linier/scatterplot#:~:text=Scatterplot%20adalah%20sebuah%20grafik%20yang,merepresentasikan%20nilai%20x%20dan%20y.

https://www.investopedia.com/terms/m/mlr.asp#:~:text=Multiple%20linear%20regression%20(MLR)%20is,uses%20just%20one%20explanatory%20variable.

https://www.w3schools.com/python/python_ml_multiple_regression.asp