UNIVERSITAS TERBUKA

Tugas ke-2

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) dan usia dalam tahun (AGE) terlampir, lakukan analisis statistik deskriptif dengan cara:

1. **Buatlah visualisasi data bivariat menggunakan Scatter Plot**
2. **Tentukan model persamaan regresi linear**
3. **Buatlah rumusan masalah asosiatif berdasarkan visual data yang ada**
4. **Berikan ulasan Anda tentang model dan hasil yang akan dicapai.**

Jawaban Tugas 2 STDA4101-24.1 (1 dan 2)

November 19, 2024

Nama: Moh Evan Alsadik | Nim: 053779188 | STDA4101

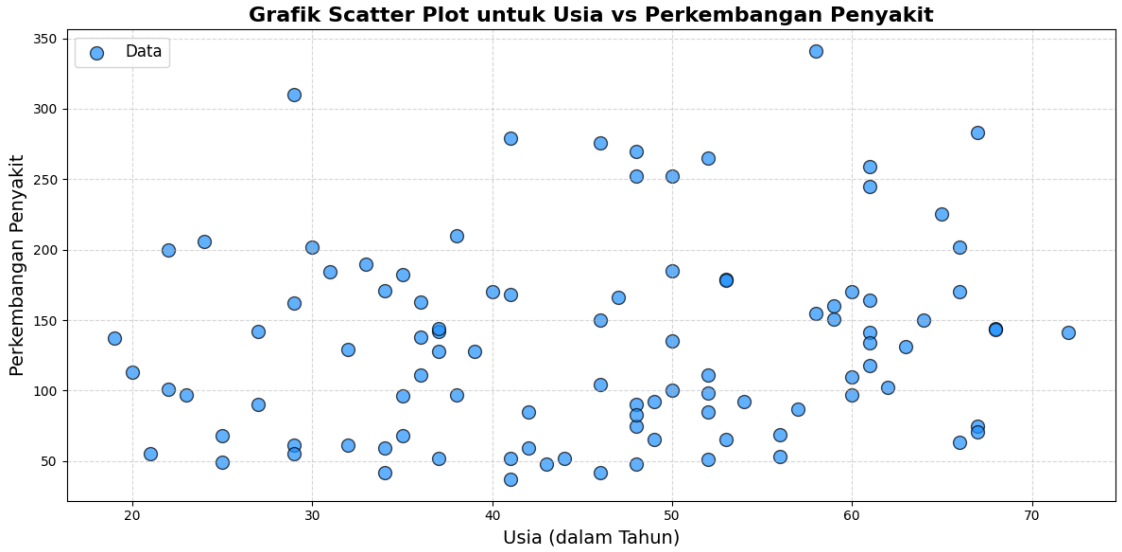
[3]:
```
from scipy.stats import linregress
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
file_tugas = "Tugas 2 STDA4101-2024.1.xlsx"
tugas_df = pd.read_excel(file_tugas)
# tugas_df.describe()
# tugas_df.info()
usia = tugas_df["AGE"]
perkembangan_penyakit = tugas_df["Y"]
print("\n")
```

Berdasarkan data ukuran kuantitatif perkembangan penyakit setelah satu tahun amatan (Y) dan usia dalam tahun (AGE) terlampir, lakukan analisis statistik deskriptif dengan cara:

0.0.1 Buatlah visualisasi data bivariat menggunakan Scatter Plot

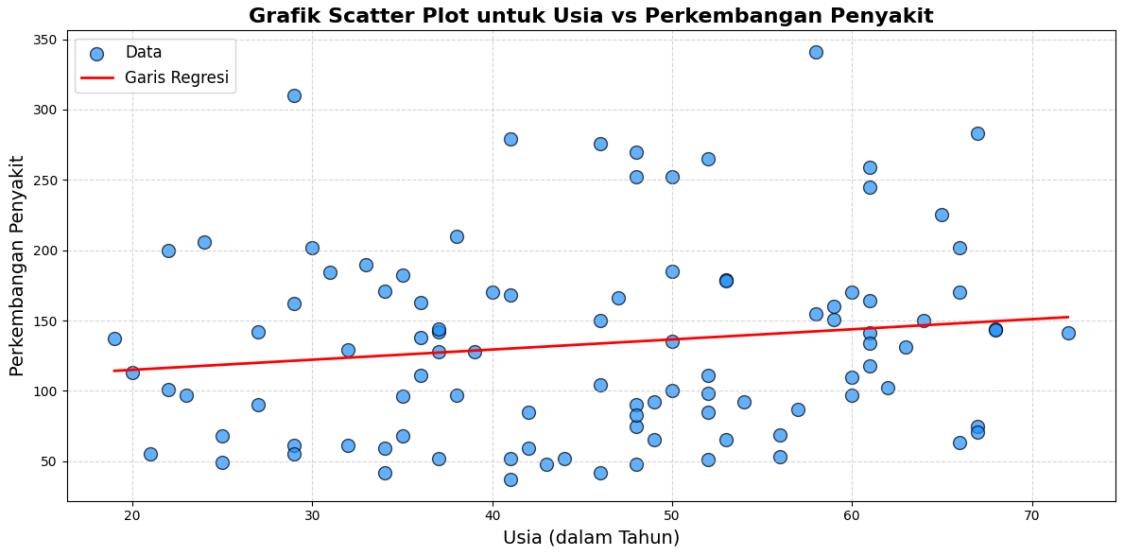
[8]:
```
# Scatter Plot
def scatter_plot():
    plt.figure(figsize=(12, 6))
    plt.scatter(
        usia,
        perkembangan_penyakit,
        color='dodgerblue',
        alpha=0.7,
        edgecolor='black',
        s=100,
        label="Data"
    )
```

```python
# Label dan Grid
def label_grid():
    # Label dan judul
    plt.title("Grafik Scatter Plot untuk Usia vs Perkembangan Penyakit",
              fontsize=16,
              weight='bold')
    plt.xlabel("Usia (dalam Tahun)", fontsize=14)
    plt.ylabel("Perkembangan Penyakit", fontsize=14)

    # Grid dan legenda
    plt.grid(True, linestyle='--', alpha=0.5)
    plt.legend(loc="upper left", fontsize=12)

scatter_plot()
label_grid()

# Tampilkan plot
plt.tight_layout()
plt.show()

print("\n")
print("\n")
print("\n")
```



0.0.2 Tentukan model persamaan regresi linear

[7]:
```
# Hitung regresi linear
slope, intercept, r_value, p_value, std_err = linregress(usia,␣
 ↪perkembangan_penyakit)

# Grafik Scatter Plot dengan Garis Regresi
scatter_plot()

# Susun data untuk garis regresi
sorted_usia = sorted(usia)
regresi_y = slope * pd.Series(sorted_usia) + intercept  # Hitung nilai Y untuk␣
 ↪setiap nilai X

# Garis regresi
plt.plot(sorted_usia, regresi_y, color="red", linewidth=2, label="Garis␣
 ↪Regresi")

label_grid()

# Tampilkan plot
plt.tight_layout()
plt.show()
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 arah hubungan␣
 ↪linier antara dua variabel (-/+)
print(f"p-value: \t\t{p_value:.4f}") # menguji signifikansi hubungan antar␣
 ↪variabel
print(f"Standard Error: \t{std_err:.2f}") # mengukur tingkat ketidakpastian␣
 ↪dalam estimasi (slope)

print("\n")
print("\n")
```




```
Persamaan Regresi Linear:
Y = 100.56 + 0.72 * X
Koefisien Korelasi (r): 0.14
p-value:                0.1572
Standard Error:         0.51
```

Soal & Pembahasan

0.0.3 Buatlah rumusan masalah asosiatif berdasarkan visual data yang ada

- Apakah model regresi linear sederhana dapat memprediksi hubungan antara usia dan perkembangan penyakit?
- Bagaimana pola hubungan yang muncul antara usia dan perkembangan penyakit berdasarkan data yang tersedia?
- Sejauh mana faktor usia (AGE) mempengaruhi perkembangan penyakit (Y) seorang pasien?

0.0.4 Berikan ulasan Anda tentang model dan hasil yang akan dicapai

Persamaan regresi linear:

$$Y = b + a X$$

ket.

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

r-value:

koefisien korelasi Pearson adalah angka yang mengukur kekuatan serta arah hubungan linier antara variabel (X dan Y).

- r = 1 : Hubungan positif sempurna, setiap ada kenaikan pada satu variabel selalu diikuti oleh kenaikan proporsional pada variabel lain.
- r = −1 : Hubungan negatif sempurna, setiap ada kenaikan pada satu variabel selalu diikuti oleh penurunan proporsional pada variabel lain.
- r = 0 : Tidak ada hubungan linier.

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:

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:

$$Y = 100.56 + 0.72X$$

keterangan:

- Ketika X = 0, nilai rata-rata yang diprediksi untuk Y adalah 100.56 (nilai intercept). Dalam data, usia nol tahun tidak pernah ada.
- Tiap kali usia (X) bertambah 1 tahun, akan dihubungkan dengan peningkatan rata-rata perkembangan penyakit (Y) sebesar 0.72.

Selanjutnya untuk:

- Nilai (r-value = 0.14) menunjukkan hubungan yang **sangat lemah** antara kedua variabel tersebut. (nilai r yang lebih dekat dengan angka 0 menunjukkan ketiadaan hubungan linear, sedangkan jika nilai r mendekati -1 atau 1 berarti keduanya memiliki hubungan).
- Nilai (p-value = 0.1572) menunjukkan **kurangnya bukti statistik** dalam menyimpulkan bahwa hubungan kedua variabel ini signifikan secara statistik (hampir 16% hubungan yang ada terjadi secara kebetulan).
- Nilai standard error = 0.51 lebih kecil dibandingkan dengan nilai slope (0.72). Hal ini menunjukkan bahwa **model lebih stabil dan hasil estimasi slope lebih dapat dipercaya**. *Namun* karena nilai p-value yang tinggi, maka sudah cukup mendukung kesimpulan bahwa hubungan ini kemungkinan besar terjadi secara kebetulan.

Kesimpulan:

Persamaan regresi yang diperoleh adalah $Y = 100.56 + 0.72XX$. Persamaan ini menunjukkan bahwa untuk setiap kenaikan usia (1 tahun) dihubungkan dengan peningkatan rata-rata perkembangan penyakit sebesar (b = 0.72). Namun, dengan nilai koefisien korelasi yang terlalu lemah (r = 0.14) serta nilai p-value yang tinggi (p = 0.1572), membuat hubungan ini tidak signifikan secara statistik. Sehingga model kurang dapat dipercaya/diandalkan dalam memprediksi perkembangan penyakit berdasarkan usia. Untuk analisis yang lebih mendalam dan akurat, diperlukan mengeksplorasi faktor (variabel) lain dan melakukan perbaikan model dengan cara memasukkan lebih banyak variabel yang relevan.

3 Hasil yang akan dicapai berdasarkan rumusan masalah sebelumnya:

1. Menjawab Rumusan Masalah 1:
   Mendapatkan informasi mengenai apakah model regresi linear dapat memprediksi hubungan antara usia dan perkembangan penyakit. Hasil yang perlu dicapai meliputi penilaian terhadap tingkat hubungan linear antara kedua variabel.

2. Menjawab Rumusan Masalah 2:

   Mengidentifikasi pola hubungan antara usia dan perkembangan penyakit berdasarkan scatter plot dan garis regresi. Hasilnya berupa grafik visual yang menggambarkan apakah hubungan itu cenderung positif, negatif, atau tidak signifikan.

3. Menjawab Rumusan Masalah 3:

   Mengukur sejauh mana faktor usia mempengaruhi perkembangan penyakit melalui nilai koefisien slope (b) dalam persamaan regresi. Hasilnya akan memberi petunjuk apakah perubahan usia memiliki dampak signifikan terhadap perkembangan penyakit atau tidak.

Sumber Referensi

https://dqlab.id/mudah-and-gampang-buat-grafik-di-excel-data-bivariat

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.linregress.html