FINAL PROJECT

DSC 2022

OLEH

IRENI LUSYANTI GILI

192400020

PROGRAM STUDI STATISTIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS PGRI ADI BUANA SURABAYA

2022

BAB I

Deskripsi Dataset

Dataset yang digunakan dalam Final Project ini adalah dataset Student_Marks (Nilai Siswa) yang bersumber dari https://www.kaggle.com. Pada dataset Student_Marks terdiri atas tiga variabel yakni dua variabel independen/predictor (X) dan satu variabel dependen/respon (Y). Berikut ini adalah rincian dari variabel dan data tersebut;

Y : Marks (nilai)

X1 : number_courses (kursus angka)

X2 : time_study (lama waktu belajar)

X1	X2	Y	X1	X2	Y
3	4,508	19,202	6	6,703	40,602
4	0,096	7,734	6	4,13	22,184
4	3,133	13,811	4	0,771	7,892
6	7,909	53,081	7	6,049	36,653
8	7,811	55,299	8	7,591	53,158
6	3,211	17,822	7	2,913	18,238
3	6,063	29,889	8	7,641	53,359
5	3,413	17,264	7	7,649	51,538
4	4,41	20,348	3	6,198	31,236
3	6,173	30,862	8	7,468	51,343
3	7,353	42,036	6	0,376	10,522
7	0,423	12,132	4	2,438	10,844
7	4,218	24,318	6	3,606	19,59
3	4,274	17,672	3	4,869	21,379
3	2,908	11,397	7	0,13	12,591
4	4,26	19,466	6	2,142	13,562
5	5,719	30,548	4	5,473	27,569
8	6,08	38,49	3	0,55	6,185
6	7,711	50,986	4	1,395	8,92
8	3,977	25,133	6	3,948	21,4
4	4,733	22,073	44	3,736	16,606
6	6,126	35,939	5	2,518	13,416
5	2,051	12,209	3	4,633	20,398

7	4,875	28,043	3	1,629	7,014
4	3,635	16,517	4	6,954	39,952
3	1,407	6,623	3	0,803	6,217
7	0,508	12,647	5	6,379	36,746
8	4,378	26,532	8	5,985	38,278
5	0,156	9,333	7	7,451	49,544
4	1,299	8,837	3	0,805	6,349
8	3,864	24,172	7	7,975	54,321
3	1,923	8,1	8	2,262	17,705
8	0,923	15,038	4	7,41	44,099
6	6,954	39,965	5	3,197	16,106
3	4,083	17,171	8	1,982	16,461
3	7,543	43,978	8	6,201	39,957
4	2,966	13,119	7	4,067	23,149
6	7,283	46,453	3	1,033	6,053
7	6,533	41,358	5	1,803	11,253
6	7,775	51,142	7	6,376	40,024
4	0,14	7,336	7	4,182	24,394
6	2,754	15,725	8	2,73	19,564
6	3,591	19,771	4	5,027	23,916
5	1,557	10,429	8	6,471	42,426
4	1,954	9,742	8	3,919	24,451
3	2,061	8,924	6	3,561	19,128
4	3,797	16,703	3	0,301	5,609
4	4,779	22,701	4	7,163	41,444
3	5,635	26,882	7	0,309	12,027
5	3,913	19,106	3	6,335	32,357

Berdasarkan data di atas dapat diketahui bahwa banyaknya sampel adalah 100 orang siswa. Data semuanya lengkap, tidak ada nilai yang hilang/missing. Dari tipe data diketahui pula bahwa jenis analisis data yang dapat digunakan adalah Analisis Regresi Linear Berganda.

BAB II

Statistik Deskriptif

Statistik deskriptif adalah metode dari pengorganisasian, penjumlahan, dan penyajian data dalam sebuah cara yang nyaman dan informatif, termasuk teknik grafik, dan teknik penghitungan. Statistik deskriptif dapat mendeskripsikan data yang sedang dianalisis, tetapi tidak boleh menarik kesimpulan apapun dari data. Statistik deskriptif dibagi menjadi dua bagian, ukuran pemusatan data dan ukuran penyebaran data.

- Ukuran Pemusatan Data

1. Mean

```
>mean = student_marks.mean()
>print('Mean:',mean)
Mean : 24.417689999999993
```

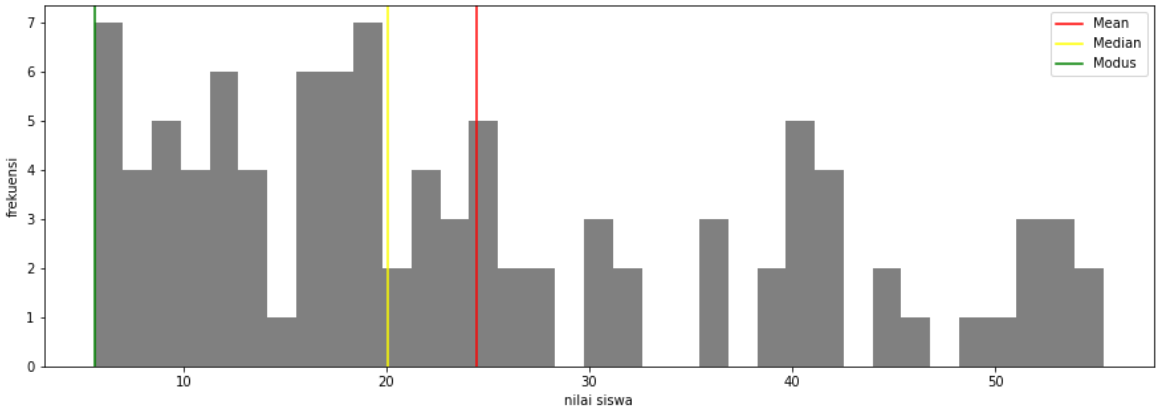
2. Median

```
>median = student_marks.median()
> print('nMedian: ',median)
nMedian:  20.0595
```

3. Modus

```
> modus = student_marks.mode()
> print('nModus: ',modus[0])
nModus:  5.609
```

4. Plot



- Ukuran Penyebaran Data

1. Nilai tertinggi

```
> maks = student_marks.max()
```

```
> print('Nilai tertinggi: ',maks)
Nilai tertinggi:  55.299
```

2. Nilai terendah

```
>minm = student_marks.min()
> print('Nilai terendah: ',minm)
nNilai terendah:  5.609
```

3. Range

```
>jarak = maks - minm
> print('nRange: ',jarak)
nRange:  49.69
```

4. Varians

```
>varians = student_marks.var()
> print('nVarians: ',varians)
nVarians:  205.23996548878785
```

5. Simpangan Baku

```
>simp_baku = student_marks.std()
> print('nSimpangan Baku: ', simp_baku)
nSimpangan Baku:  14.326198570757976
```

6. Koefisien varians

```
>koef_var = simp_baku / mean
> print('nKoefisien Variasi: ',koef_var)
nKoefisien Variasi:  0.5867139180961827
```

BAB III

Analisis Data

Regresi linier berganda merupakan model persamaan yang menjelaskan hubungan satu variabel tak bebas/ response (Y) dengan dua atau lebih variabel bebas/ predictor (X1, X2,…Xn). Tujuan dari uji regresi linier berganda adalah untuk memprediksi nilai variable tak bebas/ response (Y) apabila nilai-nilai variabel bebasnya/ predictor (X1, X2,..., Xn) diketahui. Disamping itu juga untuk dapat mengetahui bagaimanakah arah hubungan variabel tak bebas dengan variabel - variabel bebasnya.

Model regresi linear berganda dilukiskan dengan persamaan sebagai berikut:

$$\mathbf{Y = \alpha + \beta 1\ X1 + \beta 2\ X2 + \beta n\ Xn + e}$$

Keterangan:
Y = Variabel terikat atau variabel response.
X = Variabel bebas atau variabel predictor.
α = Konstanta.
β = Slope atau Koefisien estimate.

Untuk melakukan analasis regresi linear berganda menggunakan jupyter python maka langkah yang harus ditempuh adalah sebagai berikut;

```
#Import Package
>import numpy as np
>import pandas as pd
>import statsmodels
>import patsy
>import statsmodels.api as sm
>import matplotlib.pyplot as plt

#Masukin data yang telah berbentuk csv dengan memanggil datanya sesuai tempat
penyimpanannya
>data = pd.read_csv("Student_Marks.csv")
>data.head(20)
```

```
In [2]: data = pd.read_csv("Student_Marks.csv")
        data.head(20)
```

Out[2]:

	number_courses	time_study	Marks
0	3	4.508	19.202
1	4	0.096	7.734
2	4	3.133	13.811
3	6	7.909	53.018
4	8	7.811	55.299
5	6	3.211	17.822
6	3	6.063	29.889
7	5	3.413	17.264
8	4	4.410	20.348
9	3	6.173	30.862
10	3	7.353	42.036
11	7	0.423	12.132
12	7	4.218	24.318
13	3	4.274	17.672
14	3	2.908	11.397
15	4	4.260	19.466
16	5	5.719	30.548
17	8	6.080	38.490
18	6	7.711	50.986
19	8	3.977	25.133

>data.shape

```
In [3]: data.shape
Out[3]: (100, 3)
```

>print(“#Jumlah dataset = “ +str(len(data.index)))

```
In [4]: print("#Jumlah dataset = " +str(len(data.index)))
        #Jumlah dataset = 100
```

>data.describe()

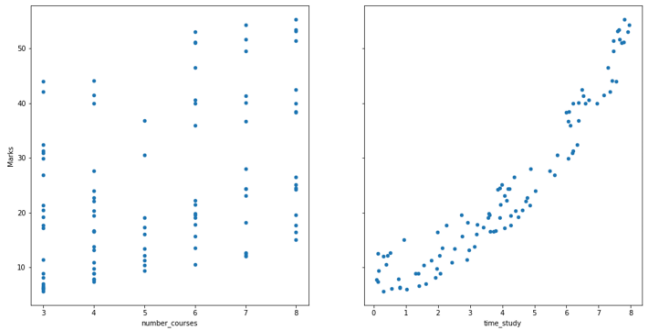
```
In [13]: data.describe()
```

Out[13]:

	number_courses	time_study	Marks
count	100.000000	100.000000	100.000000
mean	5.290000	4.077140	24.417690
std	1.799523	2.372914	14.326199
min	3.000000	0.096000	5.609000
25%	4.000000	2.058500	12.633000
50%	5.000000	4.022000	20.059500
75%	7.000000	6.179250	36.676250
max	8.000000	7.957000	55.299000

```
>fig, axs = plt.subplots(1, 2, sharey=True)
>data.plot(kind='scatter', x='number_courses', y='Marks', ax=axs[0], figsize=(16,8))
>data.plot(kind='scatter', x='time_study', y='Marks', ax=axs[1])
```

Out[14]: <matplotlib.axes._subplots.AxesSubplot at 0x18b759f1b50>




```
>feature_names=['number_courses','time_study']
>X=data[feature_names]
>X
>y=data.Marks
```

#Import model

```
>from sklearn.linear_model import LinearRegression
>from sklearn.model_selection import train_test_split
>from sklearn import metrics

>X_train, X_test, y_train, y_test = train_test_split(X,y,random_state=1)
>Linreg=LinearRegression()
>Linreg.fit(X_train,y_train)
>y_pred=Linreg.predict(X_test)
print (y_pred)
```

```
[47.55733119 18.8722928  38.61601837 20.30802166 41.89010291 39.88521011
 16.10829478 38.89536275 25.04838069 15.78792957 26.67628467 44.67171758
 47.88939848  4.85320607 25.98160445 13.48830959  8.80781493 10.91916049
 44.96276404 36.65069951 16.96460449  7.30029964 37.62884479 23.29474045
 29.10186089]
```

```
>import statsmodels.api as sm

>model=sm.OLS(y,X).fit()
>predictions=model.predict(X)
>model.summary()
```

Out[24]:

OLS Regression Results

Dep. Variable:	Marks	R-squared (uncentered):	0.979
Model:	OLS	Adj. R-squared (uncentered):	0.978
Method:	Least Squares	F-statistic:	2235.
Date:	Sat, 29 Jan 2022	Prob (F-statistic):	1.76e-82
Time:	05:03:55	Log-Likelihood:	-284.00
No. Observations:	100	AIC:	572.0
Df Residuals:	98	BIC:	577.2
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
number_courses	0.8403	0.143	5.862	0.000	0.556	1.125
time_study	5.0643	0.170	29.803	0.000	4.727	5.402

Omnibus:	9.324	Durbin-Watson:	1.710
Prob(Omnibus):	0.009	Jarque-Bera (JB):	8.550
Skew:	0.643	Prob(JB):	0.0139
Kurtosis:	2.369	Cond. No.	3.61

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

```
>X=sm.add_constant(X)
>model=sm.OLS(y,X).fit()
>model.summary()
```

Out[22]:

OLS Regression Results

Dep. Variable:	Marks	R-squared:	0.940
Model:	OLS	Adj. R-squared:	0.939
Method:	Least Squares	F-statistic:	764.8
Date:	Fri, 28 Jan 2022	Prob (F-statistic):	4.09e-60
Time:	22:46:36	Log-Likelihood:	-266.62
No. Observations:	100	AIC:	539.2
Df Residuals:	97	BIC:	547.1
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-7.4563	1.174	-6.349	0.000	-9.787	-5.125
number_courses	1.8641	0.202	9.243	0.000	1.464	2.264
time_study	5.3992	0.153	35.303	0.000	5.096	5.703

Omnibus:	29.529	Durbin-Watson:	1.978
Prob(Omnibus):	0.000	Jarque-Bera (JB):	9.956
Skew:	0.526	Prob(JB):	0.00689
Kurtosis:	1.867	Cond. No.	23.9

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Dari output tersebut didapatkan model persamaan regresi yaitu :

Y= -7,4563 + 1,8641X1 +5,3992X2

Berdasarkan model tersebut dapat disimpulkan bahwa jika kursus angka dan lama waktu belajar mendekati nol maka nilai siswa menjadi -7,4563. Sedangkan jika kursus angka naik satu satuan akan menaikan nilai satu satuan nilai siswa sebesar 1,8641 dan jika lama waktu belajar naik satu satuan maka akan menaikan nilai siswa sebesar 5,3992. Nilai R-square 0,940 atau 94% yang artinya variabel kursus angka dan lama waktu belajar secara serentak menjelaskan variabel nilai siswa.

- Uji Parsial

Uji parsial digunakan untuk menguji parameter secara parsial dengan kata lain untuk mengetahui apakah *variable independent* (X) berpengaruh secara signifikan (nyata) terhadap *variable dependent* (Y). Dari output didapat p-value (Konsatanta) sebesar 0.000 nilai (kursus angka) sebesar 0.000 dan nilai (lama waktu belajar) sebesar 0.000.

-

Hipotesis

$H_0 : \beta_i = 0$, i = 0,1,2 (Tidak terdapat pengaruh secara signifikan antara X dengan Y)

$H_1 : \beta_i \neq 0$, i = 0,1,2 (Ada pengaruh secara signifikan antara X dengan Y)

Tingkat signifikasi

$\propto$=5% = 0.05

Daerah kritis

Jika p-*value* $\leq \propto$ (0.05) $\rightarrow$ Tolak H_0

Statistik uji

P-value : = 0.000 dan = 0.000 ; $\propto$= 0,05

Keputusan

Karena nilai p-*value* untuk β_1 dan $\beta_2 < \propto$ maka tolak

Kesimpulan

Dengan tingkat kepercayaan 95% terdapat pengaruh secara signifikan antara *variable* X (kursus angka dan lama waktu belajar) dengan variabel Y(nilai siswa).\

- **Uji Normalitas**

Uji Normalitaas adalah untuk melihat apakah nilai residual terdistribusi normal atau tidak. Model regresi yang baik adalah memiliki nilai residual yang terdistribusi normal. Disini kita menggunakan nilai prob Jaque Bera (JB) dari output diatas sebesar 0,00689. Dengan hipotesis sebagai berikut:

Hipotesis

H_0 : Residual berdistribusi normal

H_1 : Residual tidak berdistribusi normal

Tingkat signifikansi

$\propto$=5% ($\propto$=0.05)

Statistik Uji

p-value = 0.00689

Daerah kritis

Tolak H_0 jika *p-value* < α

Keputusan

Karena nilai *p-value* sama dengan 0.00689, dimana nilai *p-value* >α yaitu 0.00689 > 0,05 maka gagal tolak H_0.

Kesimpulan

Jadi, dengan tingkat kepercayaan 95% dapat disimpulkan bahwa residual berdistribusi normal.

- **Uji Autokorelasi**

Dapat dilihat dari gambar output diatas bahwa uji autokorelasi menggunakan output DW(Durbin-Watson) sebesar 1,978.

Hipotesis

H_0 : Tidak terdapat Autokorelasi

H_1 : terdapat Autokorelasi

Tingkat signifikansi

$\propto$=5 ($\propto$=0,05)

Daerah kritis

Tolak H_0 : jika $0 < DW < d_l$ atau $4 - d_l < DW < 4$

Gagal tolak H_0 : jika $d_u < DW < 4 - d_u$

Tidak ada keputusan : jika $d_l < DW < d_u$ atau $4 - d_u < DW < 4 - d_l$

Statistik uji

Karena nilai *DW* 1,978 dengan $d_u < DW < 4 - d_u$ atau $0.9820 < DW < 4 - 1.5386$ maka interval daerah keputusannya yaitu $0.9820 < 1,978 < 2.4614$ sehingga dapat dikatakan bahwa Tolak H_0.

Kesimpulan

Jadi, dengan tingkat signifikansi 5% dapat disimpulkan bahwa keputusan uji adalah Tidak terdapat Autokorelasi.

- **Uji Multikolinearitas**

Uji Multikolinearitas untuk melihat ada tidaknya korelasdi antara variabel bebas dalam regresi linear berganda.

```
>from patsy import dmatrices
>from statsmodels.stats.outliers_influence import variance_inflation_factor
>lm = smf.ols(formula = 'Marks~number_courses+time_study', data = data).fit()
>y,X = dmatrices ('Marks~number_courses+time_study', data = data, return_type
="dataframe")
>vif = [variance_inflation_factor(X.values, i) for i in range (X.shape[1])]
>print(vif)
```

```
In [27]: from patsy import dmatrices
         from statsmodels.stats.outliers_influence import variance_inflation_factor
         lm = smf.ols(formula = 'Marks~number_courses+time_study', data = data).fit()
         y,X = dmatrices ('Marks~number_courses+time_study', data = data, return_type ="dataframe")
         vif = [variance_inflation_factor(X.values, i) for i in range (X.shape[1])]
         print(vif)

         [11.042136877869721, 1.0437988449385478, 1.0437988449385478]
```

Hipotesis

H_0 : VIF < 10 artinya tidak terdapat Multikolinearitas.

H1 : VIF > 10 artinya terdapat Multikolinearitas.

Taraf signifikansi

$\propto$=5% ($\propto$=0.05)

Statistik Uji

VIF :

Konstan = 11.04

Kursus Angka = 1.04

Lama waktu belajar = 1.04

Daerah Kritis

Tolak H_0 jika *VIF* > $\propto$, Tolak H_0

Keputusan

Karena nilai VIF (Kursus Angka =1.04, Lama waktu belajar = 1.04) < α maka gagal Tolak H_0

-

Kesimpulan

Jadi, dengan tingkat kepercayaan 95% dapat disimpulkan bahwa tidak terdapat Multikolinearitas.

BAB IV

Kesimpulan

Berdasarkan hasil output dari python didapatkan model yaitu Y= -7,4563 + 1,8641X1 +5,3992X2 yang mana model tersebut mendefinisikan bahwa variabel X (kursus angka dan lama waktu belajar) secara bersaman/serentak dapat menjelaskan variabel Y (nilai siswa). Selain itu dapat disimpulkan bahwa pula data tersebut memenuhi asmusi yaitu data berdistribusi normal, tidak terdapat autokorelasi dan tidak terjadi multikolinearitas.