

E-MODUL

PRAKTIKUM



MATA KULIAH
STATISTIKA REGRESI
Dengan Bahasa Pemrograman Python

Aviolla Terza Damaliana
UPN “VETERAN” JAWA TIMUR

KATA PENGANTAR

Puji dan syukur penulis panjatkan kepada Tuhan Yang Maha Esa atas selesainya penyusunan E-Modul Praktikum Mata kuliah Statistika Regresi dengan Bahasa Pemrograman Python. Tak lupa juga mengucapkan salawat serta salam semoga senantiasa tercurahkan kepada Nabi Besar Muhammad SAW, karena berkat beliau, kita mampu keluar dari kegelapan menuju jalan yang lebih terang. Kami ucapkan juga rasa terima kasih kami kepada pihak-pihak yang mendukung lancarnya penulisan e-modul praktikum ini, yaitu orang tua kami, rekan-rekan kami, dan masih banyak lagi yang tidak bisa kami sebutkan satu per satu.

E-Modul Praktikum ini dipersiapkan terutama untuk mahasiswa Program Studi Sains Data yang sedang mempelajari Statistika Regresi, karena sepanjang pengalaman penulis mengeajar mata kuliah ini mahasiswa lebih menyukai praktek dengan data dibandingkan dengan mempelajari teorinya.

Oleh karena itu, dalam e-modul ini tidak hanya berisi teori mengenai statistika regresi saja melainkan juga langkah-langkah analisisnya sehingga menjadi alternatif pegangan bagi mahasiswa dan dosen yang menempuh studi tersebut.

Penulis sadar, masih banyak kesalahan yang tentu saja jauh dari sempurna tentang buku ini. Oleh sebab itu, kami mohon agar pembaca memberi kritik dan juga saran terhadap e-modul praktikum ini agar kami dapat terus meningkatkan kualitas e-modul ini.

Demikian e-modul praktikum ini penulis susun agar dengan harapan bahwa pembaca baik mahasiswa maupun masyarakat luas dapat memahami statistika regresi tidak hanya teori namun juga langkah analisisnya menggunakan bahasa pemrograman Python. Terima kasih.

DAFTAR ISI

KATA PENGANTAR	ii
DAFTAR ISI	iii
DAFTAR TABEL	vi
DAFTAR GAMBAR	vii
PETUNJUK PENGGUNAAN.....	ix
PENDAHULUAN	1
I. LATAR BELAKANG	1
II. DESKRIPSI SINGKAT	2
III. TUJUAN PEMBELAJARAN	2
IV. DESKRIPSI MATERI PRAKTIKUM	3
BAB I PENGENALAN PYTHON	5
I. Tujuan Pembelajaran	5
II. Uraian Materi	5
1. Keterkaitan Python dengan Statistika	5
2. Google Colaboratory	6
3. Library Pada Python	7
III. Rangkuman.....	9
IV. Soal.....	10
BAB II METODE KORELASI	11
I. Tujuan Pembelajaran	11
II. Uraian Materi	11
1. Pengertian Analisis Regresi.....	11
2. Pengertian Analisis Korelasi	12
3. Perbedaan Regresi dengan Korelasi	13
III. Rangkuman.....	13
IV. Tutorial Metode	14
1. Metode Korelasi Pearson.....	14
2. Metode Korelasi Spearman	16
3. Metode Korelasi Tau Kendall	17
V. Instruksi Tugas.....	19
VI. Soal Korelasi	20
BAB III REGRESI LINIER SEDERHANA.....	21
I. Tujuan Pembelajaran	21
II. Uraian Materi	21
1. Definisi Regresi Linier Sederhana	21
2. Teori Regresi Linier Sederhana.....	22

III. Rangkuman.....	26
IV. Tutorial Metode	26
1. Studi Kasus Regresi Linier Sederhana.....	26
2. Langkah-langkah Penyelesaian.....	27
V. Instruksi Tugas.....	37
VI. Soal Regresi Linier Sederhana.....	38
BAB IV ASUMSI RESIDUAL.....	40
I. Tujuan Pembelajaran	40
II. Uraian Materi	40
1. Teori Asumsi Residual Regresi Linier.....	40
III. Rangkuman.....	48
IV. Tutorial Metode	49
1. Studi Kasus Analisis Residual	49
2. Langkah-langkah Penyelesaian.....	49
V. Instruksi Tugas.....	57
VI. Soal Asumsi Residual	57
BAB V REGRESI LINIER BERGANDA	58
I. Tujuan Pembelajaran	58
II. Uraian Materi	58
1. Teori Regresi Linier Berganda	58
2. Teori Regresi Linier Sederhana.....	59
III. Rangkuman.....	64
IV. Tutorial Metode	64
1. Studi Kasus Regresi Linier Berganda.....	64
2. Langkah-langkah Penyelesaian.....	65
V. Instruksi Tugas.....	85
VI. Soal Regresi Linier Berganda.....	85
BAB VI REGRESI DUMMY	88
I. Tujuan Pembelajaran	88
II. Uraian Materi	88
1. Teori Regresi Variabel Dummy	88
III. Rangkuman.....	91
IV. Tutorial Metode	91
1. Studi Kasus Regresi Variabel Dummy	91
2. Langkah-langkah Penyelesaian.....	93
V. Instruksi Tugas.....	110
VI. Soal Regresi Variabel Dummy	111
BAB VII REGRESI POLYNOMIAL	113
I. Tujuan Pembelajaran	113
II. Uraian Materi	113

1. Teori Regresi Variabel Polynomial.....	113
III. Rangkuman.....	114
IV. Tutorial Metode	115
1. Studi Kasus Regresi Polynomial	115
2. Langkah-langkah Penyelesaian.....	115
VI. Instruksi Tugas	122
VI. Soal Regresi Variabel Polinomial.....	122
PENUTUP	124
KUNCI JAWABAN.....	125
DAFTAR PUSTAKA	129

DAFTAR TABEL

Tabel 1.	Perbedaan Metode Korelasi dengan Regresi	13
Tabel 2.	Data Nilai Statistika Regresi dengan IPK mahasiswa	15
Tabel 3.	Data Mahasiswa Mata Kuliah Bahasa Inggris	20
Tabel 4.	Uji Statistik F.....	24
Tabel 5.	Data Pendapatan dan Pengalaman Bekerja.....	27
Tabel 6.	Data Produk Minuman di 35 Negara	38
Tabel 7.	Uji Statistik F.....	61
Tabel 8.	Data Lama Lulus dan Faktor yang Mempengaruhi	65
Tabel 9.	Data Persentase Kesempatan Mahasiswa Diterima di Pasca Sarjana	86
Tabel 10.	Data Waktu Penyelesaian Produk	92
Tabel 11.	Data Harga Jual Mobil.....	111
Tabel 12.	Data Percobaan	115
Tabel 13.	Data Ukuran Ikan Berdasarkan Usia	123

DAFTAR GAMBAR

Gambar 1.	Tampilan Awal Notebook Baru Colab	6
Gambar 2.	Scatter Plot antara IPK dengan Nilai Statistika Regresi	23
Gambar 3.	Izin Mengakses Google Drive.....	29
Gambar 4.	Memilih Akun Google Drive	29
Gambar 5.	Output Data Pendapatan	30
Gambar 6.	Output setelah nama Variabel di Ganti	30
Gambar 7.	Scatter Plot Pengalaman Bekerja vs Pendapatan.....	32
Gambar 8.	Output Model Regresi Linier Sederhana.....	33
Gambar 9.	Plot Q-Q Data ke-1.....	42
Gambar 10.	Plot Q-Q Data ke-2	43
Gambar 11.	Plot residual dengan urutan pengamatan ke-1	43
Gambar 12.	Plot Residual dengan Urutan Data ke-2.....	44
Gambar 13.	Plot Standardized Residual dengan Nilai Prediksi ke-1.....	45
Gambar 14.	Plot Standardized Residual dengan Nilai Prediksi ke-2	45
Gambar 15.	Plot Q-Q Data Residual Pendapatan	50
Gambar 16.	Plot Residual dengan Urutan Pengamatan	51
Gambar 17.	Plot Standardized Residual dengan Nilai Prediksi	52
Gambar 18.	Izin Mengakses Google Drive Data Regresi Berganda.....	67
Gambar 19.	Memilih Akun Google Drive untuk Data Regresi Berganda	67
Gambar 20.	Output Data Lama Lulus Mahasiswa	68
Gambar 21.	Mengganti Nama Variabel di Data Lama Lulusan.....	68
Gambar 22.	Scatter Plot Data IQ vs Lama Lulus.....	70
Gambar 23.	Scatter Plot Data IPK vs Lama Lulus.....	70
Gambar 24.	Output Model Regresi Linier Berganda	72
Gambar 25.	Plot Q-Q data Residual Lama Lulusan	77
Gambar 26.	Plot Residual Lama Lulusan dengan Data urutannya	78
Gambar 27.	Plot Nilai Prediksi Lama Lulusan dengan Standardized Residual.....	79
Gambar 28.	Fungsi Garis Regresi Variabel Dummy	89

Gambar 29.	Izin Mengakses Google Drive Data Regresi Berganda	94
Gambar 30.	Memilih Akun Google Drive untuk Data Regresi Berganda	94
Gambar 31.	Output Data Waktu Penyelesaian Mesin.....	95
Gambar 32.	Mengganti Nama Variabel di Data Waktu Penyelesaian Produk.....	95
Gambar 33.	Output Model Regresi Dummy.....	98
Gambar 34.	Plot Q-Q data Residual Waktu Penyelesaian Produk	102
Gambar 35.	Plot Residual Waktu Penyelesaian Produk	103
Gambar 36.	Plot Nilai Waktu Penyelesaian Produk dengan Standardized Residual	104
Gambar 37.	Izin Mengakses Google Drive Data Regresi Berganda	116
Gambar 38.	Memilih Akun Google Drive untuk Data Regresi Berganda	117
Gambar 39.	Output Data Percobaan.....	117
Gambar 40.	Mengganti Nama Variabel di Data Hasil Percobaan	118
Gambar 41.	Output Model Regresi Polynomial orde pertama	119
Gambar 42.	Output Model Regresi Polynomial orde ke-dua.....	120

PETUNJUK PENGGUNAAN

Petunjuk Bagi Mahasiswa

Untuk memahami mata kuliah ini dengan maksimal, Langkah-langkah yang perlu dilaksanakan dalam e-modul praktikum ini antara lain:

1. Baca dan pahami teori maupun langkah-langkah analisis yang ada pada setiap bab kegiatan belajar. Bila terdapat hal yang tidak dipahami, mahasiswa dapat bertanya kepada dosen
2. Kerjakan setiap tugas sesuai dengan instruksi setiap bab kegiatan pembelajaran
3. Mahasiswa dapat mengevaluasi sendiri hasil pekerjaannya dengan melihat kunci jawaban dari beberapa soal yang telah tersedia.

Petunjuk Bagi Dosen

Dalam setiap bab kegiatan pembelajaran, dosen berperan untuk:

1. Membantu mahasiswa untuk memahami e-modul praktikum dengan baik
2. Menilai dan mengevaluasi ketercapaian mahasiswa dalam mengerjakan tugas e-modul praktikum

PENDAHULUAN

I. LATAR BELAKANG

Statistika regresi merupakan salah satu mata kuliah wajib untuk mahasiswa Program Studi Sains Data UPN “Veteran” Jawa Timur. Mata kuliah ini mempelajari konsep analisis regresi linier untuk menyelesaikan permasalahan yang berkaitan dengan hubungan antara dua variabel dependen (atau variabel respon) dengan variabel independen (atau variabel prediktor) dalam berbagai bidang seperti bisnis, ekonomi, pemerintahan, ilmu sosial, bahkan hingga bidang Kesehatan. Hubungan antara dua variabel tersebut memungkinkan untuk mengetahui seberapa besar pengaruh perubahan satu unit variabel independen terhadap variabel dependennya serta dapat memprediksi variabel dependen berdasarkan variabel independen yang telah diketahui.

Mata kuliah ini tidak hanya mempelajari teori regresi linier saja namun juga mempraktekkan metode dengan data real menggunakan *software* atau pemrograman komputer. Biasanya *software* yang sering digunakan untuk menganalisis regresi linier oleh *statistician* antara lain SPSS, Minitab, E-Views, SAS, dan Stata. Namun, dikarenakan *software* statistika tersebut berbayar, para *statistician* juga menggunakan pemrograman komputer yang open source (artinya tidak berlisensi) seperti Bahasa pemrograman R dan atau Bahasa pemrograman Python.

Sebagai data analyst memiliki kewajiban untuk menguasai program komputer dalam mengolah data. Berkaitan dengan mata kuliah Statistika Regresi, maka pengguna e-modul harus bisa mengolah dan menganalisis dua variabel data yaitu independen dan dependen menggunakan program komputer dengan Bahasa pemrograman R atau Python. Kedua bahasa pemrograman tersebut saat ini sangat popular digunakan pada dunia kerja. Namun, dikarenakan pengguna e-modul tidak hanya menganalisis suatu data saja maka lebih baik membiasakan menggunakan Bahasa pemrograman Python.

Mata kuliah ini membutuhkan suatu perangkat pembelajaran yang tidak hanya mempelajari teori saja namun juga dapat melakukan praktikum metode regresi

menggunakan data dari berbagai bidang. Adapun perangkat pembelajaran yang saat ini dibutuhkan mahasiswa adalah berbentuk e-modul praktikum karena perkembangan jaman yang saat ini serba digital. Selain itu juga untuk mencapai profil lulusan Program Studi Sains Data UPN "Veteran" Jawa Timur sebagai data analyst maka praktikum mata kuliah ini menggunakan bahasa pemrograman Python. Oleh karena itu, penulis sebagai dosen pengampu mata kuliah ini menyusun e-modul praktikum statistika regresi menggunakan bahasa pemrograman Python sebagai upaya untuk meningkatkan kesempatan setiap mahasiswa menjadi data analyst saat lulus nanti.

II. DESKRIPSI SINGKAT

Dalam e-modul praktikum mata kuliah statistika regresi dengan bahasa pemrograman Python akan memuat materi-materi antara lain:

- a. Pengenalan bahasa pemrograman python
- b. Analisis regresi dan korelasi
- c. Analisis regresi linier sederhana
- d. Analisis Asumsi regresi linier
- e. Analisis regresi linier berganda,
- f. Analisis regresi dummy, dan
- g. Analisis regresi polynomial.

Dalam setiap bab materi pada e-modul praktikum ini akan memuat tujuan pembelajaran, uraian materi, rangkuman, tutorial metode, instruksi tugas, dan soal yang harus dijawab oleh pengguna e-modul. Soal yang didapat pengguna berupa data yang akan dianalisis menggunakan metode regresi dengan Bahasa pemrograman Python. Pengguna e-modul diharapkan dapat menjawab semua soal yang diberikan sebagai acuan bahwa telah mencapai tujuan pembelajaran yang diinginkan.

III. TUJUAN PEMBELAJARAN

Tujuan pembelajaran mahasiswa setelah selesai mempelajari e-modul praktikum mata kuliah statistika regresi adalah dapat melakukan pengolahan, menganalisis, membuat model regresi dari data atau informasi hasil

pengamatan, mengetahui variabel independen yang berpengaruh terhadap variabel dependen, dan dapat memprediksi variabel dependen berdasarkan model yang dibangun dengan bahasa pemrograman Python.

IV. DESKRIPSI MATERI PRAKTIKUM

Berikut adalah deskripsi singkat mengenai materi-materi yang akan ditulis pada e-modul praktikum mata kuliah Statistika Regresi dengan bahasa pemrograman Python:

a. Pengenalan bahasa pemrograman python

Pada bab ini akan dipelajari mengenai beberapa hal antara lain: 1) Keterkaitan bahasa pemrograman Python dengan metode Statistika, 2) Apa itu *Google Colaboratory*, 3) Apa saja Library pada Python

b. Perbedaan analisis regresi linier dan korelasi

Pada bab ini akan dipelajari mengenai beberapa hal antara lain: 1) Pengertian analisis regresi, 2) Pengertian analisis korelasi, 3) Perbedaan antara keduanya, dan 4) Mengetahui korelasi antara dua variabel menggunakan metode korelasi Pearson, Spearman, dan Tau Kendall dimana penyelesaiannya dengan bahasa pemrograman Python.

c. Analisis regresi linier sederhana

Pada bab ini akan dipelajari mengenai beberapa hal antara lain: 1) Pengertian analisis regresi sederhana, 2) Metode analisis regresi sederhana 3) Menganalisis dua variabel yaitu variabel dependen dan satu variabel independen menggunakan metode regresi linier sederhana dengan bahasa pemrograman Python.

d. Analisis Asumsi Residual pada regresi linier

Pada bab ini akan dipelajari mengenai beberapa hal antara lain:, 1) Teori mengenai asumsi residual dari regresi linier dan bagaimana cara memeriksanya, serta 2) Menganalisis residual dengan cara memeriksa apakah dalam model regresi asumsi residualnya terpenuhi atau tidak menggunakan bahasa pemrograman Python.

e. Analisis regresi linier berganda,

Pada sub-bab ini akan dipelajari mengenai beberapa hal antara lain: 1) Metode analisis regresi berganda, 2) Menganalisis variabel dependen dan variabel independen lebih dari satu menggunakan metode regresi linier berganda dengan bahasa pemrograman Python, dan 3) Menganalisis terpenuhinya asumsi residualnya.

f. Analisis regresi variabel dummy.

Pada bab ini akan dipelajari mengenai beberapa hal antara lain: 1) Teori mengenai regresi variabel dummy, dan 2) Menganalisis menggunakan metode regresi linier variabel dummy, dimana penyelesaiannya dengan bahasa pemrograman Python.

g. Analisis regresi variabel polynomial,

Pada sub-bab ini akan dipelajari mengenai beberapa hal antara lain: 1) Teori mengenai metode analisis regresi dengan variabel polynomial , dan 2) Menganalisis suatu data menggunakan metode regresi linier variabel polynomial dengan bahasa pemrograman Python.

BAB I

PENGENALAN PYTHON

I. Tujuan Pembelajaran

Tujuan pembelajaran mahasiswa setelah selesai mempelajari bab I mengenai python antara lain:

- a. Dapat mengetahui keterkaitan bahasa pemrograman Python dengan metode statistika,
- b. Dapat mengetahui apa itu *Google Colaboratory*, dan
- c. Dapat mengetahui library pada Python.

II. Uraian Materi

Berikut adalah uraian materi bab I mengenai Python sehingga dapat menjawab tujuan pembelajaran yang telah ditetapkan:

1. Keterkaitan Python dengan Statistika

Dalam menganalisis data menggunakan metode Statistika dapat dilakukan dengan *Software* berbayar atau dengan pemrograman komputer secara gratis. Pemrograman komputer secara gratis yang dapat digunakan salah satunya adalah bahasa pemrograman Python. Keterkaitan dengan statistika adalah bahasa pemrograman ini telah didukung oleh Library seperti Numpy, Pandas, Statmodels, Scipy, dan Scikit-Learn pada Python untuk mengolah dan menganalisis data menggunakan metode statistika. Meskipun tidak sedetail bahasa pemrograman R dalam menganalisis metode statistika, namun untuk orang yang tidak benar-benar menguasai metode statistika, bahasa pemrograman Python lebih mudah dipelajari dan digunakan. Seperti namanya, Pemrograman Python berbentuk bahasa atau kode yang diketik sesuai dengan tujuan penggunanya. Bahasa pemrograman ini relatif

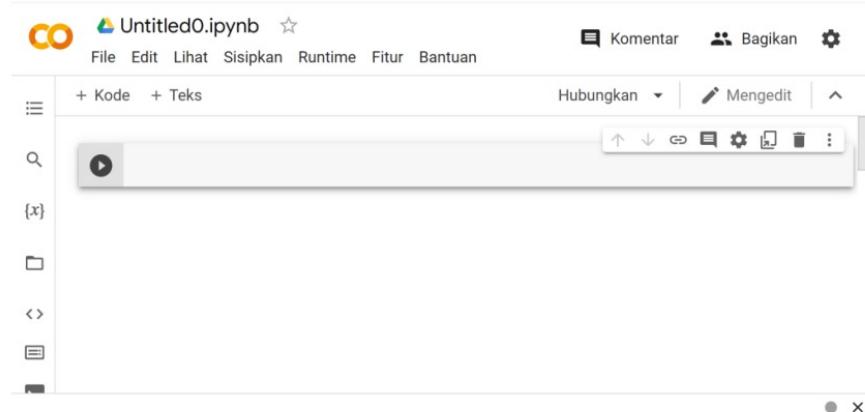
mudah dipelajari karena terfokus pada sintaks dan tata Bahasa yang sederhana.

2. Google Colaboratory

Dalam menggunakan bahasa pemrograman Python dapat dilakukan dengan 2 cara, yaitu menginstall di desktop laptop masing-masing dan menggunakan layanan secara online. Salah satu layanan dalam menggunakan bahasa pemrograman Python secara online adalah Google Colaboratory atau dikenal sebagai Google Colab. Layanan ini mereplikasi Jupiter Notebook tetapi berbasis *cloud Google*. Google Colab memungkinkan pengguna menjalankan kode Python tanpa perlu melakukan proses instalasi terlebih dahulu dan proses setup lainnya. Oleh karena itu layanan ini merupakan yang terbaik bagi mahasiswa yang ingin mengasah pengetahuan dan ketrampilan mengolah atau menganalisis data menggunakan bahasa pemrograman python.

Langkah-langkah pertama kali menggunakan Google Colab antara lain:

- a) Pengguna harus memiliki akun google untuk mengakses layanan ini karena jika tidak maka sebagian fitur dari layanan tidak dapat digunakan atau tidak berfungsi.
- b) Pergi ke alamat website <https://colab.research.google.com/>
- c) Membuat notebook baru pada menu File → Notebook baru.



Gambar 1. Tampilan Awal Notebook Baru Colab

- d) Setelah menampilkan Gambar 1, layanan untuk menggunakan bahasa pemrograman Python dapat segera digunakan.

3. Library Pada Python

Library pada dunia pemrograman komputer merupakan tempat dokumentasi kumpulan kode yang sebelumnya sudah dikompilasi. Kumpulan kode ini membuat pemrograman Python menjadi lebih sederhana dan memudahkan programmer karena tidak perlu menulis kode yang sama berulang kali untuk program yang berbeda.

Library pada Python berperan vital dalam banyak bidang antara lain bidang *machine learning*, *data science*, visualisasi data, statistika modeling, dan banyak yang lainnya. Berikut adalah macam-macam library Python yang popular dikalangan data scientist:

a) Tensorflow

Tensorflow merupakan platform end-to-end open-source untuk machine learning yang memiliki ekosistem alat, library, dan sumber daya komunitas yang fleksibel sehingga memungkinkan para peneliti mendorong dan mengembangkan teknologi atau aplikasi berbasis machine learning. Untuk memanggil library-nya, pengguna dapat menulis kode berikut ini pada Notebook Google Colab anda:

```
import tensorflow as tf
```

b) NumPy

NumPy atau Numerical Python merupakan library python yang digunakan untuk bekerja dengan array, memiliki fungsi dalam domain aljabar linier, transformasi fourier, dan matriks. NumPy ini menyediakan objek array hingga 50 kali lebih cepat daripada array dalam Python tradisional dan merupakan tempat men-generate multi dimensi yang efisien karena tipe data yang tidak jelas pun dapat didefinisi. Untuk memanggil library-nya, pengguna dapat menulis kode berikut ini pada Notebook Google Colab anda:

```
import numpy as np
```

c) Scipy

Scipy merupakan open-source library yang digunakan untuk perhitungan kebutuhan matematikan, sains, dan teknik tingkat tinggi. Terdapat paket inti penyerta dari library ini antara lain: NumPy,

Matplotlib, iPython, Sympy, dan Pandas. Mereka bekerjasama untuk menangani komputasi yang kompleks. Untuk memanggil library-nya, pengguna dapat menulis kode berikut ini pada Notebook Google Colab anda:

```
import scipy as sp
```

d) **Pandas**

Panda merupakan library open source yang penting bagi para data scientist. Library ini dipergunakan untuk machine learning yang menyediakan struktur data tingkat tinggi dan fleksibel untuk berbagai alat analisis. Selain itu juga memudahkan pengguna untuk menganalisis data, manipulasi data, dan pembersihan data. Untuk memanggil library-nya, pengguna dapat menulis kode berikut ini pada Notebook Google Colab anda:

```
import pandas as pd
```

e) **Matplotlib**

Matplotlib merupakan library open-source python yang memplotkan data dalam 2D untuk digunakan dalam analisis data. Dalam library ini anda dapat membuat plot, histogram, diagram batang, diagram lingkaran, scatter plot, grafik, dll. Untuk memanggil library-nya, pengguna dapat menulis kode berikut ini pada Notebook Google Colab anda:

```
import matplotlib.pyplot as plt
```

f) **Scikit-Learn**

Scikit-Learn merupakan library python yang digunakan untuk data kompleks. Library ini digunakan untuk menyelesaikan permasalahan machine learning dengan didukung oleh berbagai algoritma seperti regresi linier, klasifikasi, clustering, dan lain sebagainya. Selain itu, library ini bekerja sama dengan library lain seperti NumPy dengan SciPy.

g) **Statmodels**

Statmodels merupakan modul yang menyediakan fungsi untuk estimasi model, pengujian, dan eksplorasi data statistik. Dalam

memodelkan suatu data dengan metode statistic, pengguna dapat menggunakan modul ini untuk melakukan pemodelan regresi linier, Generalized Linier models, Robust Linier models, Anova, Time series analysis, metode nonparametric, dll. Modul ini dipanggil bersamaan dengan library NumPy dengan Pandas yaitu

```
import numpy as np  
import pandas as pd  
import statsmodels.api as sm
```

h) Seaborn

Seaborn adalah library pada Python untuk memvisualisasikan data berdasarkan matplotlib. Pada libarary ini akan menyuguhkan grafik statistik yang menarik dan informatif. Biasanya library ini digunakan sebagai Langkah awal dalam menganalisis untuk mengetahui *insight* dan *knowledge* dari suatu data. Untuk memanggil library-nya, pengguna dapat menulis kode berikut ini pada Notebook Google Colab anda:

```
import seaborn as sns
```

III. Rangkuman

Berikut ini adalah rangkuman materi pada Bab 1 mengenai Python:

- a) Keterkaitan dengan statistika adalah bahasa pemrograman ini telah didukung oleh Library seperti Numpy, Pandas, Statmodels, Scipy, dan Scikit-Learn pada Python untuk mengolah dan menganalisis data menggunakan metode statistika.
- b) Google Colab memungkinkan pengguna menjalankan kode Python tanpa perlu melakukan proses instalasi terlebih dahulu dan proses setup lainnya
- c) Macam-macam library Python yang popular dikalangan data scientist antara lain Tenserflow, NumPy, Scipy, Pandas, Matplotlib, Scikit-Learn, Statmodels, dan Seaborn.

IV. Soal

Untuk mengukur pemahaman pengguna modul dalam mempelajari bab 1 mengenai Python, maka pengguna wajib menjawab soal berikut ini:

1. Mengapa Python berkaitan dengan metode statistika ?

Jawaban:

2. Apakah ada cara lain menggunakan bahasa pemrograman Python selain menginstall di desktop computer ? Jika ada sebutkan dan bagaimana cara menggunakannya?

Jawaban:

3. Sebutkan dan jelaskan library Python yang dapat digunakan untuk menganalisis data dan memvisualisasikan data ?

Jawaban:

BAB II

METODE KORELASI

I. Tujuan Pembelajaran

Tujuan pembelajaran mahasiswa setelah selesai mempelajari bab II mengenai metode korelasi antara lain:

- a. Dapat mengetahui mengenai analisis regresi,
- b. Dapat mengetahui mengenai analisis korelasi,
- c. Dapat mengetahui perbedaan antara kedua metode analisis yaitu regresi dengan korelasi
- d. Dapat mengetahui cara penyelesaian korelasi dua variabel dengan bahasa pemrograman Python menggunakan metode korelasi Pearson, Spearman, dan Tau Kendall.

II. Uraian Materi

Berikut adalah uraian materi bab II sehingga dapat menjawab tujuan pembelajaran yang telah ditetapkan:

1. Pengertian Analisis Regresi

Analisis Regresi pada dasarnya merupakan serangkaian proses metode statistika yang menyelidiki hubungan antara variabel dependen (atau respon) dan variabel independen (atau prediktor). Regresi memiliki banyak aplikasi di hampir setiap bidang, termasuk teknik, ilmu fisika dan kimia, ekonomi, manajemen, ilmu kehidupan dan biologi, dan ilmu sosial. Tujuan dari penggunaan analisis regresi antara lain:

- a) Memperkirakan hubungan antara variabel dependen dan independen,
- b) Mengidentifikasi tren dalam data,
- c) Membantu dalam memprediksi nilai nyata / konstan.

- d) Menentukan faktor yang paling penting atau faktor yang paling tidak penting dari variabel independen terhadap variabel dependennya dan bagaimana masing-masing variabel saling mempengaruhi.

Contoh pengaplikasian penggunaan analisis regresi misalnya adalah missal kita ingin mengetahui apa saja faktor yang mempengaruhi pendapatan seseorang setiap bulannya. Ternyata terdapat teori yang mengatakan bahwa faktor yang mempengaruhi pendapatan seseorang setiap bulannya adalah Usia, Banyaknya jam bekerja, dan Tingkat Pendidikan. Berdasarkan hal itu kita ingin mencari tahu apakah ketiga faktor tersebut benar-benar merupakan faktor yang mempengaruhi tingkat pendapatan seseorang dengan menggunakan analisis regresi. Kita dapat mendefinisikan pendapatan seseorang setiap bulannya sebagai variabel dependen (variabel respon) dan Usia, banyaknya jam bekerja, serta tingkat pendidikan tinggi sebagai variabel independen (variabel prediktor).

2. Pengertian Analisis Korelasi

Analisis Korelasi merupakan salah satu metode statistika yang menyelidiki hubungan dua variabel. Metode ini digunakan untuk mengetahui seberapa kuat atau lemahnya hubungan antar dua variabel. Contoh pengaplikasian metode korelasi misalnya adalah kita ingin mengetahui apakah variabel pendapatan mempunyai hubungan yang kuat dengan variabel tingkat Pendidikan. Metode ini akan dilakukan sebelum melakukan analisis regresi linier untuk memperkuat dugaan adanya hubungan antara variabel independen dengan variabel dependen. Untuk menentukan apakah kedua variabel memiliki hubungan korelasi ditandai dengan nilai koefisien korelasi antara -1 hingga 1 . Macam-macam metode analisis regresi yang akan dibahas pada e-modul praktikum ini antara lain metode korelasi Pearson, metode korelasi Spearman, dan metode korelasi Tau Kendall. Metode ini akan mempunyai 5 kesimpulan yang dapat diinterpretasikan antara lain:

- a) Kedua variabel mempunyai korelasi positif yang kuat jika nilai dari koefisien korelasi ≥ 0.7 sampai dengan 1. Artinya nilai satu variabel meningkat maka akan diikuti dengan peningkatan variabel lainnya.
- b) Kedua variabel mempunyai korelasi positif yang sedang jika jika nilai koefisien korelasi < 0.7 sampai dengan ≥ 0.5
- c) Kedua variabel mempunyai korelasi negatif yang kuat jika nilai dari koefisien korelasi ≥ -0.7 sampai dengan -1 . Artinya jika nilai satu variabel meningkat maka variabel lainnya menurun.
- d) Kedua variabel mempunyai korelasi negatif yang sedang jika jika nilai koefisien korelasi < -0.7 sampai dengan ≥ -0.5
- e) Kedua variabel memiliki hubungan atau korelasi yang lemah karena memiliki nilai koefisien korelasi < 0.5 atau < -0.5 sampai dengan 0.

3. Perbedaan Regresi dengan Korelasi

Berdasarkan pengertian dari analisis regresi dengan analisis korelasi, maka berikut adalah tabel perbedaan kedua metode.

Tabel 1. Perbedaan Metode Korelasi dengan Regresi

Metode Korelasi	Metode Regresi
Hubungan antar dua variabel	Variabel satu berpengaruh terhadap variabel yang lain
Menyatakan pergerakan dua variabel	Variabel satu merupakan variabel penyebab dan satunya adalah variabel akibat
Kedua variabel dapat ditukar	Kedua variabel tidak dapat ditukar
Kedua variabel digambarkan oleh scatter plot setiap titik	Kedua variabel digambarkan oleh adanya garis linier pada scatter plot

III. Rangkuman

Berikut ini adalah rangkuman materi pada Bab II:

- a) Analisis Regresi menyelidiki hubungan antara variabel dependen (atau respon) dan variabel independen (atau prediktor).
- b) Metode korelasi digunakan untuk mengetahui seberapa kuat atau lemahnya hubungan antar dua variabel.
- c) Perbedaan analisis regresi dengan korelasi ada 4 yaitu metode korelasi menyatakan hubungan namun metode regresi menyatakan satu variabel berpengaruh terhadap variabel lainnya, metode korelasi menyatakan

pergerakan dua variabel sedangkan metode regresi menyatakan sebab akibat, variabel metode korelasi data ditukar namun sebaliknya dengan metode regresi, dan jika metode korelasi digambarkan pada setiap titik sedangkan regresi pada suatu garis linier.

IV. Tutorial Metode

Berikut adalah tutorial materi bab II mengenai metode korelasi sehingga dapat menjawab tujuan pembelajaran yang ke-4 yaitu dapat mengetahui cara penyelesaian masalah dengan bahasa pemrograman Python menggunakan metode korelasi Pearson, Spearman, dan Tau Kendall.

1. Metode Korelasi Pearson

Metode korelasi Pearson merupakan salah satu metode korelasi antara dua variabel berjenis data kontinu. Metode ini merupakan salah satu metode statistika parametrik, sehingga kedua variabel memiliki asumsi random dan berdistribusi normal. Selain itu syarat lainnya adalah jumlah data yang sama pada kedua variabel. Misalkan jika variabel satu berjumlah 100 data maka variabel kedua juga harus berjumlah 100 data juga. Berikut adalah rumus dari metode korelasi Pearson.

$$\rho = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1)$$

dimana:

x adalah variabel yang pertama dan y adalah variabel yang kedua

Berikut adalah langkah-langkah penyelesaian metode korelasi pearson dengan menggunakan bahasa pemrograman Python pada permasalahan dua variabel yaitu variabel nilai mata kuliah statistika regresi dengan IPK yang diperoleh mahasiswa. Dataset kedua variabel dapat dilihat pada tabel berikut ini.

Tabel 2. Data Nilai Statistika Regresi dengan IPK mahasiswa

Variabel	Nilai									
Nilai Mata Kuliah	90	80	85	65	70	65	95	82	83	75
IPK	3.85	3.72	3.82	3.2	3.4	3.1	3.9	3.75	3.76	3.49

- 1) Memasukkan data tersebut ke dalam notebook Google Colab dengan kode bahasa pemrograman yaitu

```
#Menggunakan library pandas untuk membuat data frame
import pandas as pd
Nilai_Mata_Kuliah = [90, 80, 85, 65, 70, 65, 95, 82, 83, 75]
IPK = [3.85, 3.72, 3.82, 3.2, 3.4, 3.1, 3.9, 3.75, 3.76, 3.49]
df = pd.DataFrame({'X':Nilai_Mata_Kuliah, 'Y': IPK})
data = df[['X', 'Y']]
print(data)
```

Kode berikut akan menghasilkan output Python yaitu

	X	Y
0	90	3.85
1	80	3.72
2	85	3.82
3	65	3.20
4	70	3.40
5	65	3.10
6	95	3.90
7	82	3.75
8	83	3.76
9	75	3.49

- 2) Menghitung nilai korelasi Pearson

```
#Mengitung nilai korelasi pearson
from scipy.stats import pearsonr
# Convert dataframe into series
list1 = data['X']
list2 = data['Y']
corr, _ = pearsonr(list1, list2)
print(corr)
```

Kode berikut akan menghasilkan output Python yaitu

0.9604136736352115

3) Interpretasi

Berdasarkan nilai korelasi pearson sebesar 0.9604136736352115 atau dibulatkan 4 desimal dibelakang menjadi 0.9604 maka dapat disimpulkan bahwa variabel nilai mata kuliah statistika regresi dengan variabel IPK memiliki hubungan korelasi positif yang kuat karena memiliki nilai koefisien korelasi mendekati 1.

2. Metode Korelasi Spearman

Metode korelasi Spearman merupakan salah satu metode korelasi antara dua variabel dimana kedua variabel minimal memiliki jenis data ordinal. Oleh karena itu, metode ini termasuk dalam metode statistika nonparametric yang tidak memiliki asumsi distribusi normal. Syarat lainnya adalah jumlah kedua data juga harus sama. Berikut adalah rumus dari metode korelasi Spearman.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (2)$$

dimana:

d_i adalah perbedaan rank antara nilai variabel pertama dan kedua

n adalah banyaknya data

Berikut adalah langkah-langkah penyelesaian metode korelasi Spearman dengan menggunakan bahasa pemrograman Python pada permasalahan yang dapat dilihat pada Tabel 2.

- 1) Memasukkan data tersebut ke dalam notebook Google Colab dengan kode bahasa pemrograman yaitu

```
#Menggunakan library pandas untuk membuat data frame
import pandas as pd
Nilai_Mata_Kuliah = [90, 80, 85, 65, 70, 65, 95, 82, 83, 75]
IPK = [3.85, 3.72, 3.82, 3.2, 3.4, 3.1, 3.9, 3.75, 3.76, 3.49]
df = pd.DataFrame({'X':Nilai_Mata_Kuliah, 'Y': IPK})
data = df[['X', 'Y']]
```

```
print(data)
```

Kode berikut akan menghasilkan output Python yaitu

	X	Y
0	90	3.85
1	80	3.72
2	85	3.82
3	65	3.20
4	70	3.40
5	65	3.10
6	95	3.90
7	82	3.75
8	83	3.76
9	75	3.49

2) Menghitung nilai korelasi Spearman

```
#Mengitung nilai korelasi spearman
from scipy.stats import spearmanr
# Convert dataframe into series
list1 = data['X']
list2 = data['Y']
corr, _ = spearmanr(list1, list2)
print(corr)
```

Kode berikut akan menghasilkan output Python yaitu

```
0.9969650916353059
```

3) Interpretasi

Berdasarkan nilai korelasi spearman sebesar 0.9969650916353059 atau dibulatkan 4 desimal dibelakang menjadi 0.9970 maka dapat disimpulkan bahwa variabel nilai mata kuliah statistika regresi dengan variabel IPK memiliki hubungan korelasi positif yang kuat karena memiliki nilai koefisien korelasi mendekati 1.

3. Metode Korelasi Tau Kendall

Metode korelasi Tau Kendall merupakan salah satu metode korelasi antara dua variabel yang memiliki minimal data berjenis ordinal. Sama seperti metode korelasi Spearman, metode ini termasuk dalam statistika nonparametrik yang tidak memiliki asumsi data berdistribusi normal dan

jumlah datanya sama. Berikut adalah rumus dari metode korelasi Tau Kendall.

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \quad (3)$$

dimana:

τ adalah koefisien korelasi Tau Kendall

n_c adalah banyaknya *concordant*

n_d adalah banyaknya *disconcordant*

Berikut adalah langkah-langkah penyelesaian metode korelasi Tau Kendall dengan menggunakan bahasa pemrograman Python pada permasalahan yang dapat dilihat pada Tabel 2.

- 1) Memasukkan data tersebut ke dalam notebook Google Colab dengan kode bahasa pemrograman yaitu

```
#Menggunakan library pandas untuk membuat data frame
import pandas as pd
Nilai_Mata_Kuliah = [90, 80, 85, 65, 70, 65, 95, 82, 83, 75]
IPK = [3.85, 3.72, 3.82, 3.2, 3.4, 3.1, 3.9, 3.75, 3.76, 3.49]
df = pd.DataFrame({'X':Nilai_Mata_Kuliah, 'Y': IPK})
data = df[['X', 'Y']]
print(data)
```

Kode berikut akan menghasilkan output Python yaitu

	X	Y
0	90	3.85
1	80	3.72
2	85	3.82
3	65	3.20
4	70	3.40
5	65	3.10
6	95	3.90
7	82	3.75
8	83	3.76
9	75	3.49

- 2) Menghitung nilai korelasi Tau Kendall

```
#Mengitung nilai korelasi Tau Kendall
from scipy.stats import kendalltau
```

```
# Convert dataframe into series
list1 = data['X']
list2 = data['Y']
corr, _ = kendalltau(list1, list2)
print('Koefisien Tau Kendall: %.5f' % corr)
```

Kode berikut akan menghasilkan output Python yaitu

```
Koefisien Tau Kendall: 0.98883
```

3) Interpretasi

Berdasarkan nilai korelasi spearman sebesar 0.98883 atau dibulatkan 4 desimal dibelakang menjadi 0.9883 maka dapat disimpulkan bahwa variabel nilai mata kuliah statistika regresi dengan variabel IPK memiliki hubungan korelasi positif yang kuat karena memiliki nilai koefisien korelasi mendekati 1.

V. Instruksi Tugas

Setelah memahami langkah-langkah penyelesaian permasalahan data menggunakan metode korelasi dengan bahasa pemrograman Python, pengguna e-modul praktikum dapat menyelesaikan tugas dengan instruksi sebagai berikut

- 1) Melakukan analisis data dengan bahasa pemrograman Python memanfaatkan layanan Google Colab
- 2) Hasil jawaban soal diketik menggunakan File Word ukuran A4 dengan format sebagai berikut:
 - a. Cover Laporan memuat Judul, Identitas pengguna, dan, Identitas Institusi
 - b. Daftar Isi, jika ada tabel dan gambar maka harus ada Daftar Tabel dan Daftar Gambar
 - c. Bab I Pendahuluan memuat latar belakang, tujuan, dan manfaat
 - d. Bab II Tinjauan Pustaka memuat teori dari metode
 - e. Bab III Metodologi data memuat dataset yang digunakan serta Langkah-langkah penyelesaian

- f. Bab IV Hasil analisis memuat hasil penyelesaian berdasarkan soal permasalahan
 - g. Bab V Kesimpulan memuat hasil kesimpulan berdasarkan bab IV. Kesimpulan tidak boleh hasil dari copy paste
 - h. Bab VI Daftar Pustaka
- 3) Laporan dikumpulkan berbentuk Pdf.
- 4) File Notebook Google Colab hasil pemrograman Python juga ikut dikumpulkan dengan laporan
- 5) Laporan dan File Notebook Google Colab dimasukan kedalam satu folder dimana folder tersebut diberi nama identitas pengguna.

VI. Soal Korelasi

Dosen bahasa Inggris di suatu universitas ingin mengetahui apakah nilai Toefl IBT pada awal semester 20 mahasiswa berkorelasi dengan nilai akhir mata kuliah bahasa Inggris yang didapat. Data Nilai Toefl dengan nilai akhir mata kuliah bahasa inggris disajikan pada Tabel berikut ini

Tabel 3. Data Mahasiswa Mata Kuliah Bahasa Inggris

Data ke-	Nilai Toefl IBT	Nilai Akhir Kelas
1	118	90
2	107	74
3	104	71
4	110	77
5	103	71
6	115	85
7	109	76
8	101	70
9	102	72
10	108	76
11	106	74
12	111	78
13	112	79
14	109	77
15	104	72
16	105	73
17	107	75
18	106	73
19	110	78
20	102	70

Kerjakan Data tersebut dengan menggunakan metode korelasi pearson, spearman, dan Tau Kendall sesuai dengan instruksi tugas diatas.

BAB III

REGRESI LINIER SEDERHANA

I. Tujuan Pembelajaran

Tujuan pembelajaran mahasiswa setelah selesai mempelajari bab III mengenai metode Regresi Linier Sederhana antara lain:

- a. Dapat mengetahui definisi regresi linier sederhana
- b. Dapat mengetahui teori metode regresi linier sederhana
- c. Dapat mengetahui cara penyelesaian metode regresi linier sederhana dengan bahasa pemrograman Python

II. Uraian Materi

Berikut adalah uraian materi bab III sehingga dapat menjawab tujuan pembelajaran yang telah ditetapkan:

1. Definisi Regresi Linier Sederhana

Regresi Linier Sederhana merupakan metode regresi linier yang menyelidiki hubungan antara satu variabel dependen (atau respon) dan satu variabel independen (atau prediktor). Sebelum melakukan analisis regresi linier, kedua variabel independen dan dependen dibuat scatter plot untuk mengetahui apakah keduanya memiliki hubungan yang linier atau tidak. Berikut adalah model regresi linier sederhana.

$$y = \beta_0 + \beta_1 x_1 + \varepsilon \quad (4)$$

dimana

β_0 adalah intersep

β_1 adalah koefisien parameter variabel dependen

x_1 adalah variabel independen atau variabel prediktor

y adalah variabel dependen atau variabel respon

ε adalah error random yang memiliki asumsi $\varepsilon \sim iidn(0, \sigma^2)$ dimana iidn merupakan singkatan dari independen, identik, dan berdistribusi normal.

Berikut adalah langkah-langkah dalam menganalisis data menggunakan metode regresi linier sederhana

- a) Membuat Scatter plot untuk mengetahui hubungan linier antara satu variabel dependen dengan satu variabel independennya.
- b) Mengetahui seberapa kuat hubungan korelasi antara variabel dependen dengan satu variabel dependennya menggunakan metode korelasi Pearson.
- c) Melakukan estimasi untuk mendapatkan nilai estimator $\hat{\beta}_0$ dan $\hat{\beta}_1$
- d) Melakukan pengujian hipotesis Uji F untuk melihat apakah model yang didapat layak untuk digunakan
- e) Melakukan perhitungan R-squared untuk melihat seberapa besar kesesuaian model yang didapatkan
- f) Melakukan pengujian hipotesis Uji t untuk melihat apakah variabel independen berpengaruh terhadap variabel dependennya.
- g) Membuat kesimpulan dan interpretasi model
- h) Menghitung nilai prediksi dari model yang didapatkan

2. Teori Regresi Linier Sederhana

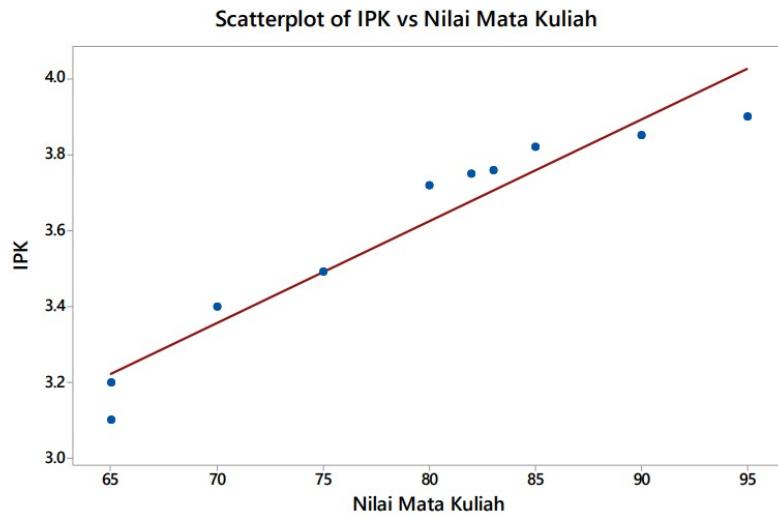
Berikut ini adalah penjelasan mengenai teori regresi linier sederhana yang lebih detail.

A. Hubungan Linier Variabel Independen dengan Dependend

Asumsi yang paling dasar dalam analisis regresi adalah adanya hubungan linier antara variabel independen dengan variabel dependennya yang dapat dilihat menggunakan *scatter plot*. Berikut ini adalah contoh *scatter plot* antara variabel independen dengan variabel dependen dari data yang diambil pada Tabel 2.

Gambar 2 terlihat bahwa titik-titik data berada disekitar garis regresinya sehingga dapat disimpulkan bahwa Nilai Statistika Regresi memiliki hubungan yang linier dengan IPK mahasiswa.

Dalam analisis regresi linier sederhana, karena mempunyai satu variabel independen maka hanya menggambarkan satu buah *Scatter Plot*.



Gambar 2. Scatter Plot antara IPK dengan Nilai Statistika Regresi

B. Estimasi Model Regresi Linier Sederhana

Diketahui bahwa model regresi linier sederhana dengan pengamatan $i = 1, 2, \dots, n$ adalah sebagai berikut.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (5)$$

dimana

x_i adalah variabel independen atau variabel prediktor

y_i adalah variabel dependen atau variabel respon

ε_i adalah error random yang memiliki asumsi $\varepsilon_i \sim iidn(0, \sigma^2)$

Untuk mendapatkan estimator $\hat{\beta}_0$ dan $\hat{\beta}_1$ dilakukan estimasi metode OLS dengan mengubah persamaan (5) menjadi

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i \quad (6)$$

Kriteria metode OLS adalah

$$\min(S(\beta_0, \beta_1)) = \min\left(\sum_{i=1}^n \varepsilon_i^2\right) = \min\left(\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2\right) \quad (7)$$

Berdasarkan kriteria metode estimasi OLS diatas dengan menggunakan perhitungan penuruan persamaan (7) yang akan disamadengankan 0 (nol) maka hasil dari estimator $\hat{\beta}_0$ dan $\hat{\beta}_1$ adalah

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_{1i} - \left(\sum_{i=1}^n y_i \right) \left(\sum_{i=1}^n x_{1i} \right)}{\sum_{i=1}^n x_{1i}^2 - \frac{\left(\sum_{i=1}^n x_{1i} \right)^2}{n}} \quad (8)$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^n y_i}{n} - \hat{\beta}_1 \left(\frac{\sum_{i=1}^n x_{1i}}{n} \right) \quad (9)$$

Setelah mendapatkan estimator $\hat{\beta}_0$ dan $\hat{\beta}_1$ maka model regresi hasil estimasi yang didapat yaitu

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} \quad (10)$$

Oleh karena itu didapat residual yang merupakan selisih antara nilai aktual dengan nilai prediksi berdasarkan variabel dependennya. Berikut ini adalah persamaan residual yang didapat,

$$e_i = y_i - \hat{y}_i \quad (11)$$

C. Pengujian Hipotesis Uji F

Uji F pada analisis regresi linier sederhana digunakan untuk melihat apakah model yang didapat layak untuk digunakan. Berikut adalah Langkah-langkah dalam melakukan Uji Hipotesis pada Uji F

- Menentukan hipotesis null dengan hipotesis alternatifnya

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

- Menentukan level of significance atau α sebesar 5% atau 0.05
- Menentukan Uji Statistik F

Tabel 4. Uji Statistik F

	SS	df	MS	F
Regresi	SS_{Reg}	1	$MS_{Reg} = \frac{SS_{Reg}}{1}$	
Residual	SS_{Res}	$n - 2$	$MS_{Res} = \frac{SS_{Res}}{n - 2}$	$F_{hit} = \frac{MS_{reg}}{MS_{res}}$
Total	SS_{Tot}	$n - 1$		

4. Menentukan titik kritis pengujian

Dengan α sebesar 0,05, H_0 ditolak jika $F_{hit} > F_{tabel}$ dimana $F_{tabel} = F_{\alpha,1,n-2}$ atau p-value < 0,05.

5. Menentukan kesimpulan hasil pengujian

D. Nilai R-Squared

Perhitungan R-squared digunakan untuk melihat seberapa besar kesesuaian model yang didapatkan. Perhitungan ini akan melihat seberapa besar pengaruh variabel independen terhadap variabel dependennya. Berikut adalah rumus dari R-Squared

$$R^2 = \frac{SS_{Reg}}{SS_{Tot}} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{SS_{Res}}{SS_{Tot}} \quad (12)$$

Nilai R-Squared antara 0 sampai dengan 1, merupakan proporsi dari total variansi yang dijelaskan oleh model regresi.

E. Pengujian Hipotesis Uji T

Uji T pada analisis regresi linier sederhana digunakan untuk mengetahui apakah variabel independen berpengaruh terhadap variabel dependennya. Berikut adalah Langkah-langkah dalam melakukan Uji Hipotesis pada Uji T

1. Menentukan hipotesis null dengan hipotesis alternatifnya

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

2. Menentukan level of significance atau α sebesar 5% atau 0.05

3. Menentukan Uji Statistik T-hitung

$$t_{hit} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \quad (13)$$

dimana

$se(\hat{\beta}_1)$ adalah estimasi standar error

4. Menentukan titik kritis pengujian

Dengan α sebesar 0.05, H_0 ditolak jika $|t_{hit}| > t_{tabel}$ dimana

$$t_{tabel} = t_{\frac{\alpha}{2}, n-2}$$

5. Menentukan kesimpulan hasil pengujian

III. Rangkuman

Berikut ini adalah rangkuman materi pada Bab II:

- a) Regresi Linier Sederhana merupakan metode regresi linier yang menyelidiki hubungan antara satu variabel dependen (atau respon) dan satu variabel independen (atau prediktor). Metode korelasi digunakan untuk mengetahui seberapa kuat atau lemahnya hubungan antar dua variabel.
- b) Berikut adalah langkah-langkah analisis regresi sederhana antara lain: membuat scatter plot variabel dependen dengan independen, menghitung nilai korelasi variabel dependen dengan independen, mendapatkan estimator $\hat{\beta}_0$ dan $\hat{\beta}_1$, melakukan Uji F, mendapatkan R-Squared, dan melakukan Uji T, membuat kesimpulan dan interpretasi model, dan menghitung nilai prediksi dari model yang didapatkan.

IV. Tutorial Metode

Berikut adalah tutorial metode materi bab III yakni langkah-langkah analisis regresi kinier sederhana sehingga dapat menjawab tujuan pembelajaran yang ke-3 yaitu dapat mengetahui cara penyelesaian metode regresi linier sederhana dengan bahasa pemrograman Python.

1. Studi Kasus Regresi Linier Sederhana

Terdapat suatu penelitian sebelumnya yang menyatakan bahwa pendapatan seseorang dipengaruhi oleh banyaknya pengalaman bekerja. Oleh karena itu seseorang peneliti ingin memprediksi berapa pendapatan seseorang jika mempunyai pengalaman bekerja selama 10 Tahun dan ingin mengetahui apakah benar bahwa pengalaman bekerja

mempengaruhi pendapatan seseorang. Untuk menjawab hal tersebut, peneliti menggunakan data yang dapat dilihat pada Tabel 5. Peneliti akan menggunakan metode regresi linier sederhana.

Tabel 5. Data Pendapatan dan Pengalaman Bekerja

Data ke-1	Pengalaman Bekerja (Tahun)	Pendapatan (Juta)
1	1.1	3934300
2	1.3	4620500
3	1.5	3773100
4	2	4352500
5	2.2	3989100
6	2.9	5664200
7	3	6015000
8	3.2	5444500
9	3.2	6444500
10	3.7	5718900
11	3.9	6321800
12	4	5579400
13	4	5695700
14	4.1	5708100
15	4.5	6111100
16	4.9	6793800
17	5.1	6602900
18	5.3	8308800
19	5.9	8136300
20	6	9394000
21	6.8	9173800
22	7.1	9827300
23	7.9	10130200
24	8.2	11381200
25	8.7	10943100
26	9	10558200
27	9.5	11696900
28	9.6	11263500
29	10.3	12239100
30	10.5	12187200

2. Langkah-langkah Penyelesaian

Berikut adalah-langkah-langkah analisis regresi linier sederhana untuk menyelesaikan permasalahan pada studi kasus data Tabel 5 menggunakan bahasa pemrograman Python.

A. Menentukan Variabel Independen dan Variabel Dependend

Berdasarkan studi kasus permasalahan peneliti yang ingin memprediksi berapa pendapatan seseorang jika mempunyai pengalaman bekerja selama 10 Tahun dan ingin mengetahui apakah benar bahwa pengalaman bekerja mempengaruhi pendapatan seseorang, maka dapat diketahui bahwa

Variabel independen (x) adalah **Pengalaman Bekerja (Tahun)**, sedangkan variabel dependen (y) adalah **Pendapatan (Juta)**.

B. Memasukkan Data ke Notebook Google Colab

Dikarenakan jumlah data yang lebih dari 20, tidak disarankan untuk mengetik secara langsung pada Notebook Google Colab. Hal ini karena akan memakan waktu yang lama dan tidak efisien. Oleh karena itu data Tabel 5 disimpan terlebih dahulu ke bentuk File csv yang Bernama : **Data_Pendapatan**

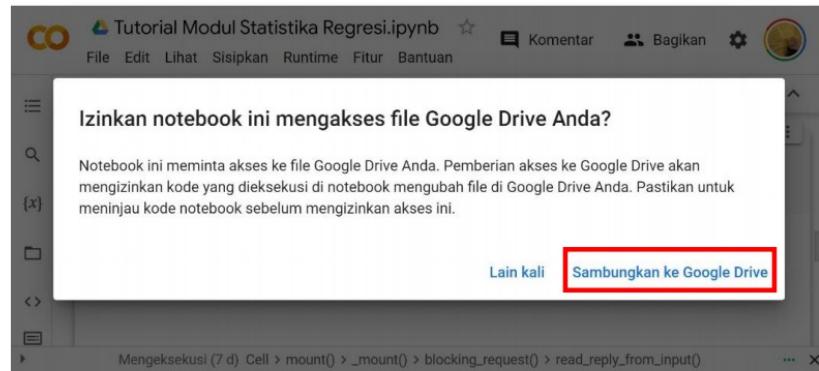
Pada e-modul Praktikum Statistika Regresi ini, dalam memanggil data bentuk File csv terlebih dahulu disimpan di Folder Google Drive masing-masing pengguna e-modul. Penulis menyimpan File csv pada Folder yang bernama : **Dataset eModul**.

Berikut adalah kode pemrograman python pada notebook Google Colab untuk memasukan data ke Notebook Google Colab

- 1) Menghubungkan Google Drive dengan Notebook Google Colab

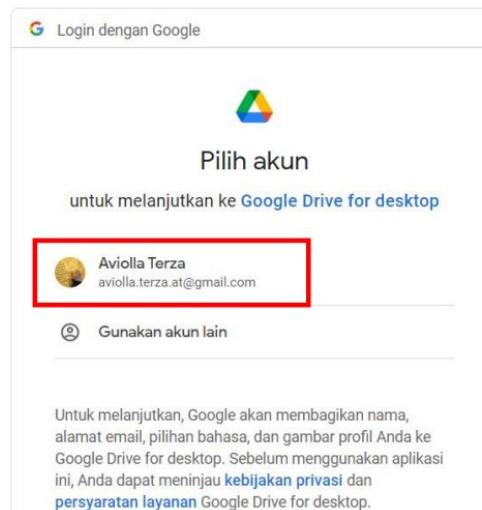
```
#CONNECT GOOGLE DRIVE
from google.colab import drive
drive.mount('/content/drive')
```

Setelah me-running kode diatas akan muncul Gambar sebagai berikut dan klik **Sambungkan ke Google Drive**



Gambar 3. Izin Mengakses Google Drive

Kemudian akan muncul Gambar berikut untuk memilih akun Google Drive tempat meyimpan dataset kita: **Klik Akun → Izinkan**



Gambar 4. Memilih Akun Google Drive

Setelah itu dataset di Google Drive sudah dapat terbaca oleh Notebook Google Colab yang ditandai dengan kode : `Mounted at /content/drive`

2) Membuat dataframe data menggunakan library Pandas

Setelah data terhubung dari Google Drive ke Notebook Google Colab, langkah selanjutnya adalah menyusun dataset kita menjadi data frame dengan library Pandas. Kode pemrograman python-nya adalah

```
# Memanggil dataset
import pandas as pd
df = pd.read_csv("drive/MyDrive/Dataset eModul/Data_Pendapatan.csv")
df.head()
```

Kode diatas akan menghasilkan Output yaitu

	Data ke-	Bekerja_Tahun	Pendapatan_Juta
0	1	1.1	3934300
1	2	1.3	4620500
2	3	1.5	3773100
3	4	2.0	4352500
4	5	2.2	3989100

Gambar 5. Output Data Pendapatan

Kolom pertama dari Gambar 5 diatas yaitu **Data ke-** akan dihitung menjadi variabel, maka harus dihilangkan dan karena nama kolom Panjang maka akan diganti **Pengalaman Bekerja (Tahun)** menjadi **x** dan **Pendapatan (Juta)** menjadi **y**. Kode pemrogramannya adalah

```
#Hilangkan kolom pertama dan mengganti nama variabel
df.drop('Data ke-', axis=1, inplace=True)
df.rename(columns={'Bekerja_Tahun':'x', 'Pendapatan_Juta':'y'}, inplace=True)
df.head()
```

Kode diatas akan menghasilkan output yaitu

	x	y
0	1.1	3934300
1	1.3	4620500
2	1.5	3773100
3	2.0	4352500
4	2.2	3989100

Gambar 6. Output setelah nama Variabel di Ganti

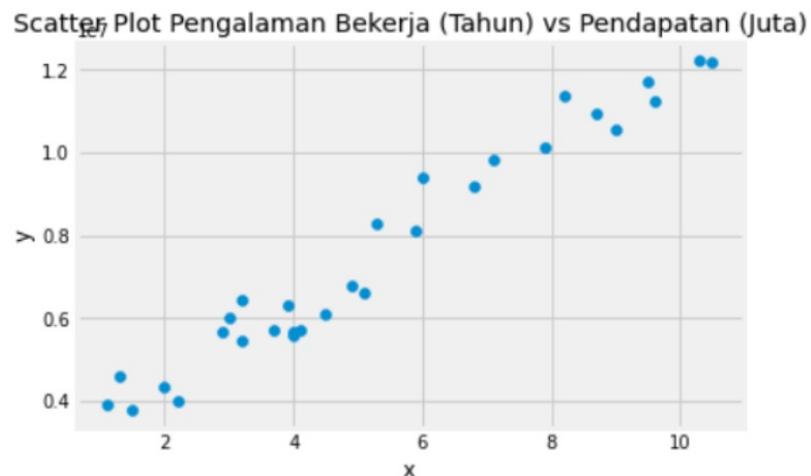
C. Mengetahui Hubungan Linier Variabel Dependen dan Independen

Untuk mengetahui apakah variabel x dan variabel y mempunyai hubungan yang linier atau tidak dapat dilihat dari Plot *Scatter Plot*. Kode pemrograman python untuk membuat plot *Scatter Plot* yaitu

```
# Library untuk memunculkan Plot
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
plt.style.use('fivethirtyeight')
import warnings
warnings.filterwarnings('ignore')
%matplotlib inline

# Untuk memunculkan Scatter Plot
plt.scatter(df['x'], df['y'])
plt.xlabel('x')
plt.ylabel('y')
plt.title('Scatter Plot Pengalaman Bekerja (Tahun) vs Pendapatan (Juta)')
plt.show()
```

Output dari Plot *Scatter plot* variabel independen dengan dependennya adalah



Gambar 7. Scatter Plot Pengalaman Bekerja vs Pendapatan

Berdasarkan Gambar 7 terlihat bahwa titik-titik data membentuk garis linier sehingga dapat disimpulkan bahwa pengalaman bekerja memiliki hubungan yang linier positif dengan pendapatan. Dari Gambar diduga bahwa tanda koefisien parameter akan bertanda positif.

D. Mengetahui Seberapa Kuat Hubungan Variabel Dependen dan Independen Menggunakan Metode Korelasi

Setelah mengetahui bahwa variabel x dan variabel y mempunyai hubungan yang linier, langkah selanjutnya adalah mengetahui seberapa kuat hubungannya dengan menggunakan metode korelasi Pearson. Berikut adalah kode pemrograman python yang digunakan

```
#Mengitung nilai korelasi pearson
from scipy.stats import pearsonr
# Convert dataframe into series
list1 = df['x']
list2 = df['y']
corr, _ = pearsonr(list1, list2)
print('Koefisien Pearson: %.5f' % corr)
```

Hasil output dari kode diatas adalah

Koefisien Pearson: 0.97824

Berdasarkan nilai korelasi Pearson sebesar 0.97824 maka dapat disimpulkan bahwa variabel pengalaman kerja dengan variabel pendapatan memiliki hubungan korelasi positif yang kuat karena memiliki nilai koefisien korelasi mendekati 1.

E. Membentuk Model Regresi Linier Sederhana

Langkah selanjutnya memodelkan dataset dengan menggunakan analisis regresi sederhana. Kode pemrograman python untuk memodelkan data dengan regresi linier sederhana adalah sebagai berikut.

```
#Memodelkan dengan Regresi Linier Sederhana
import numpy as np
import statsmodels.api as sm
x = df[['x']]
y = df['y']
x = sm.add_constant(x)
model = sm.OLS(y, x).fit()
print_model = model.summary()
print(print_model)
```

Hasil output dari Kode pemrograman Python diatas adalah

OLS Regression Results						
	Dep. Variable:	y	R-squared:	0.957		
Model:		OLS	Adj. R-squared:	0.955		
Method:	Least Squares		F-statistic:	622.5		
Date:	Tue, 15 Nov 2022		Prob (F-statistic):	1.14e-20		
Time:	02:57:17		Log-Likelihood:	-439.60		
No. Observations:	30		AIC:	883.2		
Df Residuals:	28		BIC:	886.0		
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	2.579e+06	2.27e+05	11.347	0.000	2.11e+06	3.04e+06
x	9.45e+05	3.79e+04	24.950	0.000	8.67e+05	1.02e+06
Omnibus:		2.140	Durbin-Watson:		1.648	
Prob(Omnibus):		0.343	Jarque-Bera (JB):		1.569	
Skew:		0.363	Prob(JB):		0.456	
Kurtosis:		2.147	Cond. No.		13.2	

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

Gambar 8. Output Model Regresi Linier Sederhana

Berdasarkan Gambar 8 didapatkan model regresi linier sederhana adalah

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \\ \hat{y} = (2579000) + (945000)x \quad (13)$$

Dari Persamaan (13), kita dapat menghasilkan nilai prediksi dan residualnya. Kode pemrograman python untuk menghasilkan nilai prediksi \hat{y} adalah

```
prediksi = model.predict(x)
print(prediksi.head())
```

Output hasil kode pemrograman diatas antara lain

```
0    3.618716e+06
1    3.807715e+06
2    3.996714e+06
3    4.469212e+06
4    4.658212e+06
dtype: float64
```

Adapun kode pemrograman python untuk menghasilkan nilai residual e_i adalah

```
residual=model.resid
print(residual.head())
```

Output hasil kode pemrograman diatas antara lain

```
0    315584.124773
1    812784.878344
2   -223614.368085
3   -116712.484158
4   -669111.730587
dtype: float64
```

F. Melakukan Uji F

Setelah mengetahui model regresi yang didapat, langkah selanjutnya adalah melihat apakah model yang didapat layak untuk digunakan. Untuk itu diperlukan suatu Pengujian Hipotesis Uji F dengan langkah-langkah sebagai berikut.

1. Menentukan hipotesis null dengan hipotesis alternatifnya

$H_0 : \beta_1 = 0$ (model tidak layak)

$H_1 : \beta_1 \neq 0$ (model layak)

2. Menentukan level of significance atau α sebesar 5% atau 0.05

3. Menentukan Uji Statistik F

Pada Gambar 8, didapatkan bahwa Nilai F-statistic yang didapat adalah 622.5 atau p-value yang didapat sebesar 1.14e-20

4. Menentukan titik kritis pengujian

Dengan α sebesar 0,05, H_0 ditolak jika $F_{hit} > F_{tabel}$ dimana $F_{tabel} = F_{0.05,1,28} = 4.19597$ atau p-value < 0,05.

5. Menentukan kesimpulan hasil pengujian

Dengan α sebesar 0,05, nilai $F_{hit} = 622.5$ lebih besar dibandingkan $F_{tabel} = F_{0.05,1,28} = 4.19597$ dan juga nilai p-value yang didapatkan sebesar 0.00 yang kurang dari 0,05. Berdasarkan hal itu maka dapat disimpulkan bahwa H_0 ditolak yang artinya bahwa model layak untuk digunakan

G. Mendapatkan Nilai R-Squared

Langkah selanjutnya adalah mengetahui nilai proporsi dari total varians yang dijelaskan oleh model regresi atau dikenal sebagai nilai R-Squared. Nilai ini untuk melihat seberapa besar variabel dependen dapat dijelasakan oleh variabel independennya. Berdasarkan Gambar 8, didapatkan nilai R-Squared sebesar 0.957 yang artinya bahwa 95,7 % pendapatan seseorang dipengaruhi oleh pengalaman bekerja dan sisanya 4,3% lainnya dipengaruhi oleh variabel lainnya yang tidak diketahui.

H. Melakukan Uji T

Langkah selanjutnya adalah melihat apakah variabel independen berpengaruh secara signifikan atau tidak terhadap variabel dependennya. Untuk itu diperlukan suatu Pengujian Hipotesis Uji T dengan langkah-langkah sebagai berikut.

- a. Menentukan hipotesis null dengan hipotesis alternatifnya

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

b. Menentukan level of significance atau α sebesar 5% atau 0.05

c. Menentukan Uji Statistik t-hitung

Pada Gambar 8, didapatkan bahwa Nilai t-statistic yang didapat adalah 24.950 atau p-value yang didapat sebesar 0.000

d. Menentukan titik kritis pengujian

Dengan α sebesar 0.05, H_0 ditolak jika $|t_{hit}| > t_{tabel}$ dimana $t_{tabel} = t_{\frac{\alpha}{2}, n-2} = t_{0.025; 28} = 2.048$

e. Menentukan kesimpulan hasil pengujian

Dengan α sebesar 0.05, nilai $t_{hit} = 24.950$ lebih besar dibandingkan $t_{tabel} = 2.048$ dan juga nilai p-value yang didapatkan sebesar 0.00 yang kurang dari 0.05. Berdasarkan hal itu maka dapat disimpulkan bahwa H_0 ditolak yang artinya bahwa variabel pengalaman bekerja berpengaruh signifikan terhadap pendapatan seseorang.

I. Menginterpretasikan Model

Berikut adalah model regresi yang didapatkan setelah selesai melakukan pengujian hipotesis

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{y} = (2579000) + (945000)x$$

Model diatas dapat diinterpretasikan bahwa setiap pertambahan satu satuan pengalaman bekerja akan meningkatkan pendapatan seseorang sebesar Rp 945.000 . Estimator $\hat{\beta}_0$ tidak diinterpretasikan karena tidak ada nilai variabel x yang sama dengan 0

J. Memberikan Hasil Kesimpulan

Tujuan penelitian menggunakan metode regresi linier ini adalah ingin memprediksi berapa pendapatan seseorang jika mempunyai

pengalaman bekerja selama 10 Tahun. Berikut adalah model yang didapatkan:

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x \\ \hat{y} &= (2579000) + (945000)x \\ \hat{y} &= (2579000) + (945000)(10) \\ \hat{y} &= 12029000\end{aligned}$$

Sehingga dapat disimpulkan bahwa jika seseorang mempunyai pengalaman bekerja selama 10 tahun maka pendapatan diprediksi sebesar Rp 12.029.000

Selain itu juga berdasarkan Uji T dapat disimpulkan bahwa pengalaman bekerja dapat mempengaruhi pendapatan seseorang.

V. Instruksi Tugas

Setelah memahami langkah-langkah penyelesaian permasalahan data menggunakan metode regresi linier sederhana dengan bahasa pemrograman Python, pengguna e-modul praktikum dapat menyelesaikan tugas dengan instruksi sebagai berikut

- 1) Melakukan analisis data dengan bahasa pemrograman Python memanfaatkan layanan Google Colab
- 2) Hasil jawaban soal diketik menggunakan File Word ukuran A4 dengan format sebagai berikut:
 - a. Cover Laporan memuat Judul, Identitas pengguna, dan, Identitas Institusi
 - b. Daftar Isi, jika ada tabel dan gambar maka harus ada Daftar Tabel dan Daftar Gambar
 - c. Bab I Pendahuluan memuat latar belakang, tujuan, dan manfaat
 - d. Bab II Tinjauan Pustaka memuat teori dari metode
 - e. Bab III Metodologi data memuat dataset yang digunakan serta Langkah-langkah penyelesaian
 - f. Bab IV Hasil analisis memuat hasil penyelesaian berdasarkan soal permasalahan

- g. Bab V Kesimpulan memuat hasil kesimpulan berdasarkan bab IV.
Kesimpulan tidak boleh hasil dari copy paste
 - h. Bab VI Daftar Pustaka
- 3) Laporan dikumpulkan berbentuk Pdf.
 - 4) File Notebook Google Colab hasil pemrograman Python juga ikut dikumpulkan dengan laporan
 - 5) Laporan dan File Notebook Google Colab dimasukan kedalam satu folder dimana folder tersebut diberi nama identitas pengguna.

VI. Soal Regresi Linier Sederhana

Perusahaan A mempunyai produk minuman yang telah dijual di 35 negara. Masing-masing negara, perusahaan tersebut mengiklankan produknya di satu stasiun televisi (tv). Oleh karenanya, dalam setahun data biaya iklan masing-masing stasiun televisi dan banyaknya produk yang telah terjual di masing-masing negara didapatkan. Data 35 negara yang memuat jumlah produk terjual dan biaya iklan stasiun televisi dapat dilihat pada Tabel 6. Perusahaan ingin mengetahui apakah biaya iklan di statisun televisi berpengaruh terhadap jumlah produk yang terjual. Selain itu perusahaan ingin memprediksi jumlah produk yang terjual apabila alokasi biaya iklan di televisi salah satu negara dalam setahun direncakan sebesar \$ 500 thousand.

Tabel 6. Data Produk Minuman di 35 Negara

Data ke-	Biaya iklan (\$ thousand)	Jumlah Produk terjual (million)
1	587	564
2	781	571
3	937	583
4	988	593
5	441	531
6	451	533
7	498	537
8	508	544
9	526	547
10	530	548
11	392	522
12	539	552
13	536	550
14	741	570
15	733	569
16	566	560
17	578	560
18	511	544
19	511	546

Data ke-	Biaya iklan (\$ thousand)	Jumlah Produk terjual (million)
20	796	571
21	806	571
22	297	512
23	302	513
24	487	534
25	928	582
26	408	525
27	686	567
28	347	517
29	454	533
30	495	536
31	392	519
32	821	581
33	634	566
34	252	498
35	261	511

Kerjakan Data tersebut berdasarkan tujuan dari perusahaan A dan dalam menganalisis, langkah-langkah harus dilakukan secara lengkap dan jelas.

BAB IV

ASUMSI RESIDUAL

I. Tujuan Pembelajaran

Tujuan pembelajaran mahasiswa setelah selesai mempelajari bab IV mengenai Asumsi residual regresi linier antara lain:

- a. Dapat mengetahui mengenai teori dari asumsi residual dari regresi linier
- b. Dapat melakukan pengecekan apakah residual memenuhi asumsi regresi linier yang telah ditetapkan menggunakan bahasa pemrograman Python.

II. Uraian Materi

Berikut adalah uraian materi bab IV sehingga dapat menjawab tujuan pembelajaran yang telah ditetapkan:

1. Teori Asumsi Residual Regresi Linier

Dalam analisis regresi linier, terdapat asumsi utama yang harus terpenuhi antara lain:

- a. Adanya hubungan linier antara variabel dependen y dan variabel independen x , atau minimal mendekati
- b. $Error \epsilon$ mempunyai mean 0 (null)
- c. $Error \epsilon$ mempunyai varians σ^2 yang konstan
- d. $Error \epsilon$ tidak saling berkorelasi
- e. $Error \epsilon$ berdistribusi normal

Asumsi yang menyatakan bahwa $Error \epsilon$ tidak saling berkorelasi dan berdistribusi normal secara bersama-sama mengimplikasikan bahwa error dari model regresi merupakan variabel random yang saling independen. Selain itu secara terpisah asumsi $error$ berdistribusi normal dibutuhkan untuk melakukan pengujian hipotesis maupun estimasi interval. Pelanggaran terhadap asumsi dapat menghasilkan model regresi yang tidak stabil yang artinya bahwa kemungkinan mendapatkan

koefisien yang berlawanan dari seharusnya. Akibatnya model tersebut tidak layak untuk digunakan.

A. Analisis Residual

Residual dapat dilihat pada persamaan (11) dimana merupakan deviasi atau penyimpangan antara data yang sebenarnya dengan data prediksi yang didapatkan model. Residual juga dapat mengukur variabilitas variabel dependen yang tidak dapat dijelaskan oleh model regresi. Analisis residual artinya bahwa peneliti melakukan pengecekan terhadap asumsi yang harusnya terpenuhi. Secara umum asumsi residual yang harus terpenuhi antara lain:

1. Residual berdistribusi normal

Residual berdistribusi normal diperlukan sebagai asumsi untuk melakukan estimasi parameter, Uji statistik F, Uji T, dan confidence interval pada analisis regresi linier. Jika asumsi ini tidak terpenuhi maka model tidak dapat menjelaskan hubungan dari data.

2. Residual tidak mengalami autokorelasi

Residual tidak mengalami autokorelasi artinya bahwa pengamatan residual saling independen satu dengan yang lain. Akurasi prediksi yang rendah merupakan indikasi bahwa residual mengalami autokorelasi. Jika terjadi autokorelasi berarti ada pengaruh waktu dalam modelnya, sehingga kurang tepat untuk menganalisis menggunakan regresi linier.

3. Residual tidak mengalami heteroskedastisitas

Residual tidak mengalami heteroskedastisitas artinya bahwa variansi residual konstan. Terjadinya heteroskedastisitas kemungkinan besar karena adanya outlier yang ekstrem. Ketika penyimpangan ini terjadi maka menyebabkan estimasi parameter tidak menghasilkan variansi yang minimum, sehingga deviasi antara prediksi dengan nilai aktualnya sangat jauh.

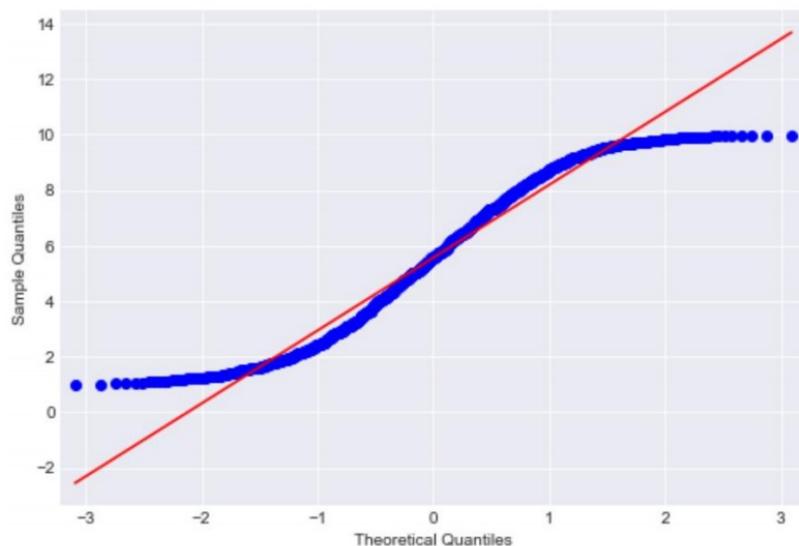
Analisis residual merupakan cara efektif untuk menemukan kekurangan dari model regresi yang kita dapat. Adapun cara yang efektif untuk menyelidiki seberapa cocok model regresi dan untuk memeriksa asumsi dapat dilakukan dengan membuat plot residual. Selain itu, juga bisa menggunakan metode pengujian statistika untuk memeriksa asumsi residual.

B. Plot Residual

Berikut adalah beberapa plot residual untuk memeriksa apakah asumsinya sudah terpenuhi atau tidak.

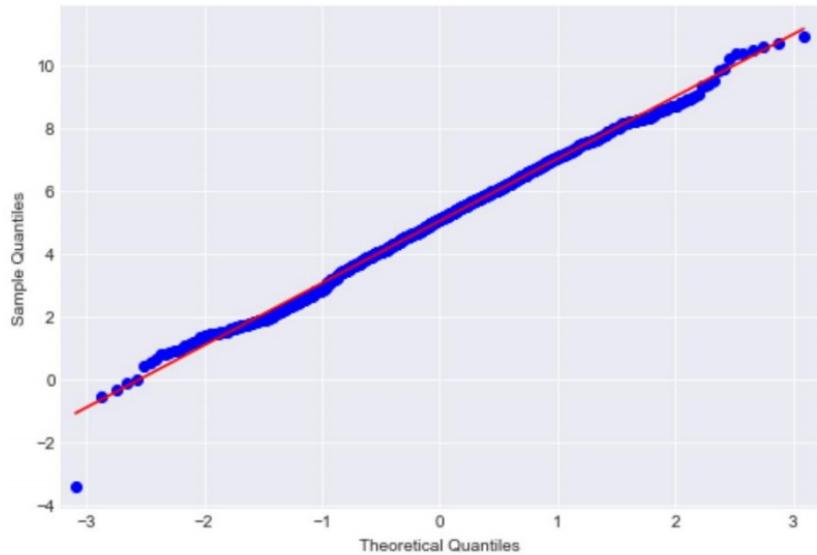
1. Asumsi residual berdistribusi normal

Plot untuk memeriksa apakah residual berdistribusi normal atau tidak menggunakan Plot Q-Q. Berikut adalah hasil dari Plot Q-Q dari suatu residual



Gambar 9. Plot Q-Q Data ke-1

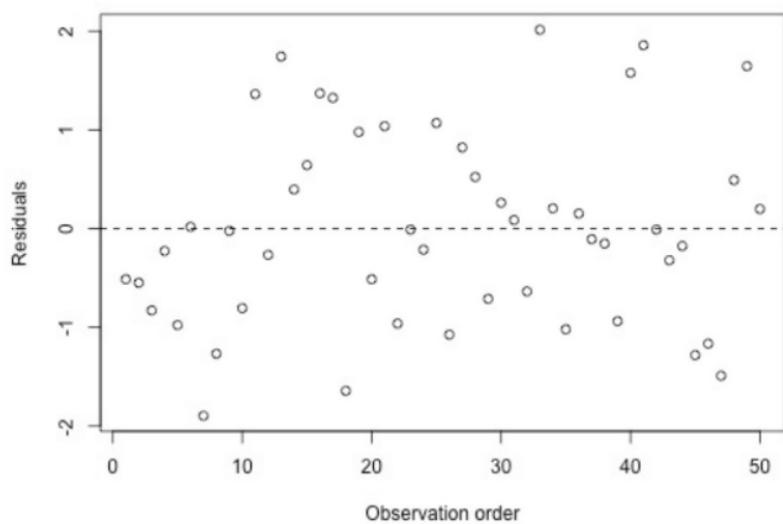
Gambar 9 menunjukkan bahwa distribusi residual sedikit menjauh dari garis linier sehingga dapat dikatakan bahwa data residual tidak berdistribusi normal. Adapun Gambar 10 menunjukkan bahwa residual berdistribusi normal karena sebagian distribusi data residual berada di garis linier.



Gambar 10. Plot Q-Q Data ke-2

2. Asumsi residual tidak terjadi autokorelasi

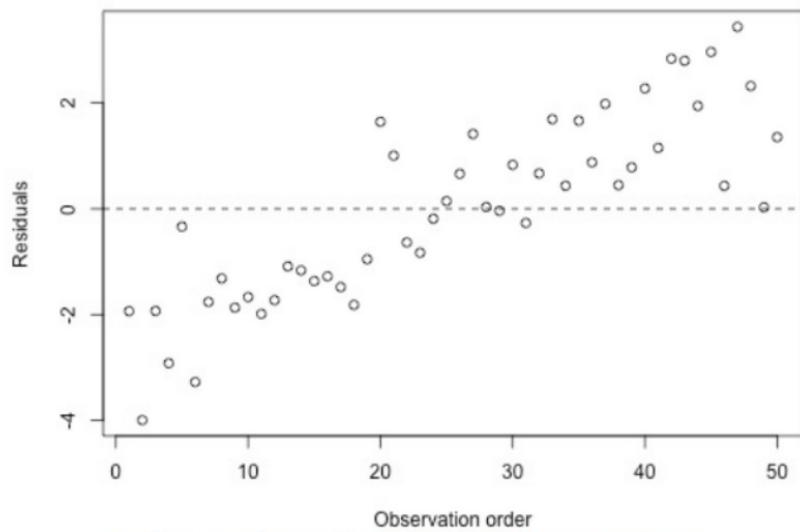
Cara untuk memeriksa apakah residual terjadi autokorelasi atau tidak yaitu dengan membuat plot antara urutan pengamatan dengan nilai residualnya. Berikut adalah beberapa contoh plot untuk menggambarkan apakah residual terjadi autokorelasi atau tidak.



Gambar 11. Plot residual dengan urutan pengamatan ke-1

Gambar 11 menunjukkan bahwa titik-titik data residual menyebar secara random dan tidak membentuk pola apapun,

sehingga dapat dikatakan asumsi residual tidak terjadi autokorelasi terpenuhi.



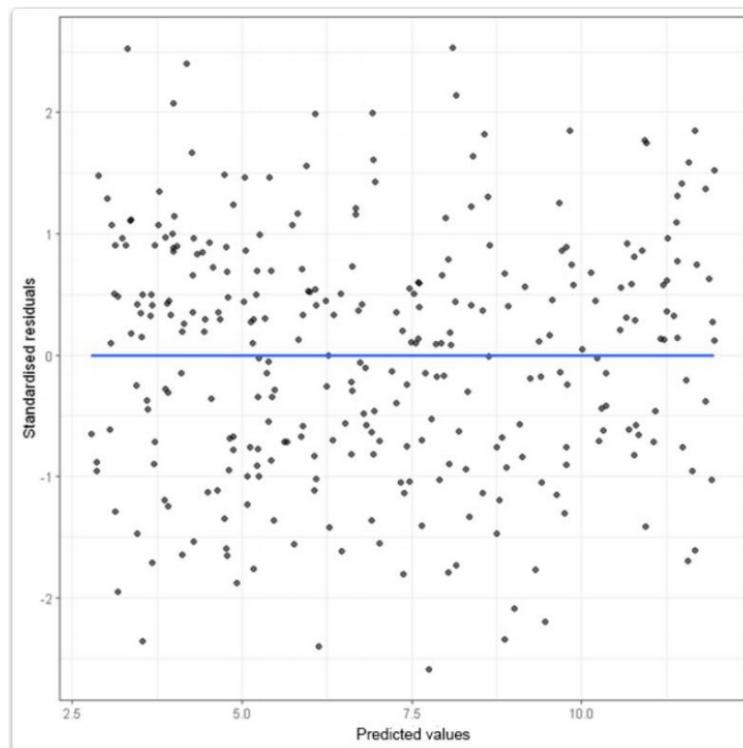
Gambar 12. Plot Residual dengan Urutan Data ke-2

Gambar 12 menunjukkan bahwa titik-titik data residual membentuk pola trend meningkat, sehingga dapat dikatakan bahwa telah terjadi autokorelasi pada pengamatan residualnya.

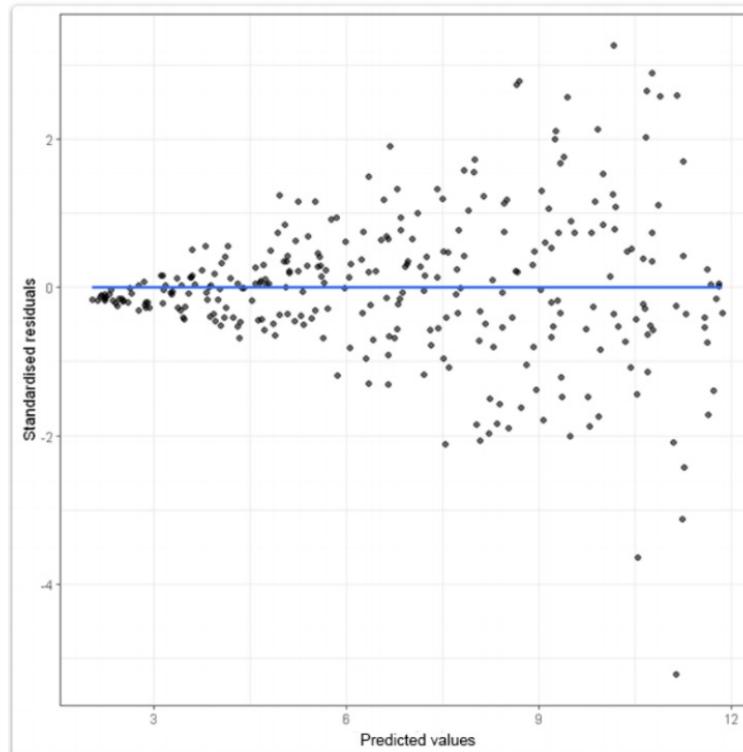
3. Asumsi residual tidak terjadi heteroskedastisitas

Cara untuk memeriksa apakah residual terjadi heteroskedastisitas atau tidak yaitu dengan membuat plot antara standardized residual dengan nilai prediksinya. Berikut adalah beberapa contoh plot untuk menggambarkan apakah residual terjadi heteroskedastisitas atau tidak.

Gambar 13 menunjukkan bahwa data menyebar secara random. Berdasarkan hal itu maka memenuhi asumsi residual tidak terjadi heteroskedastisitas yang artinya varians residualnya konstan. Adapun Gambar 14 menunjukkan bahwa semakin meningkat nilai prediksinya akan meningkatkan nilai varians residual. Hal ini berarti telah terjadi heteroskedastisitas yang artinya varians errornya tidak konstan



Gambar 13. Plot Standardized Residual dengan Nilai Prediksi ke-1



Gambar 14. Plot Standardized Residual dengan Nilai Prediksi ke-2

C. Uji Statistika Untuk Memeriksa Asumsi Residual

Memeriksa asumsi residual menggunakan plot dapat menghasilkan kesimpulan yang multitafsir antar peneliti, oleh karenanya untuk membuktikan tafsiran secara plot tersebut maka harus dibuktikan dengan menggunakan pengujian secara statistika. Berikut adalah metode-metode statistika yang digunakan untuk memeriksa asumsi residual dalam regresi linier.

- 1) Pengujian Jarque Bera untuk memeriksa asumsi residual berdistribusi normal dengan langkah-langkah sebagai berikut:

- a. Menentukan hipotesis null dengan hipotesis alternatifnya

$$H_0 : \text{Populasi berdistribusi normal}$$

$$H_1 : \text{Populasi tidak berdistribusi normal}$$

- b. Menentukan level of significance atau α sebesar 5% atau 0.05

- c. Menentukan Statistik Uji

$$JB = n \left[\frac{skewness^2}{6} + \frac{(Kurtosis - 3)^2}{24} \right] \quad (14)$$

dimana

JB adalah statistik uji Jarque Bera

$$skewness = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{3/2}}$$

$$kurtosis = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^2}$$

- d. Menentukan titik kritis pengujian

Dengan α sebesar 0.05, H_0 ditolak jika $JB > \chi^2_{(\alpha,2)}$ dimana

$$\chi^2_{(\alpha,2)} = \chi^2_{(0.05,2)} = 5.99 \text{ atau } p\text{-value} < 0.05$$

- e. Menentukan kesimpulan hasil pengujian

2) Pengujian Durbin Watson untuk memeriksa apakah residual memenuhi asumsi tidak terjadi autokorelasi. Langkah-langkah pengujian Durbin Watson adalah sebagai berikut.

a. Menentukan hipotesis null dengan hipotesis alternatifnya

$$H_0 : \text{Tidak terjadi Autokorelasi}$$

$$H_1 : \text{Terjadi Autokorelasi}$$

b. Menentukan level of significance atau α sebesar 5% atau 0.05

c. Menentukan Statistik Uji

$$d_{hit} = \frac{\sum_{t=1}^{T-1} e_t e_{t+1}}{\sum_{t=1}^T e_t^2} \quad (15)$$

dimana

d_{hit} adalah nilai statistik uji Durbin Watson

e_t adalah residual pada data ke- t .

d. Menentukan titik kritis pengujian

Dengan α sebesar 0.05, maka kemungkinan kesimpulan pengujian ini adalah

$$H_0 \text{ ditolak jika } d_{hit} < d_L,$$

$$H_0 \text{ gagal ditolak jika } d_{hit} > d_U$$

Tidak dapat menarik kesimpulan $d_L \leq d \leq d_U$

e. Menentukan kesimpulan hasil pengujian

3) Pengujian Breush-Pagan untuk memeriksa apakah residual memenuhi asumsi tidak terjadi heteroskedastisitas. Langkah-langkah pengujian Breush-Pagan adalah sebagai berikut.

a. Menentukan hipotesis null dengan hipotesis alternatifnya

$$H_0 : \text{Tidak terjadi Heteroskedastisitas}$$

$$H_1 : \text{Terjadi Heteroskedastisitas}$$

b. Menentukan level of significance atau α sebesar 5% atau 0.05

c. Menentukan Statistik Uji

Berikut beberapa Langkah untuk mendapatkan statistik uji Breush-Pagan:

1. Melakukan pemodelan regresi linier untuk mendapatkan prediksi dari variabel dependen
2. Dari prediksi yang didapat, akan mendapatkan pula residual dan residual kuadrat SS_{res} dari model
3. Melakukan pemodelan regresi linier kembali menggunakan SS_{res} sebagai variabel dependen yang baru
4. Dari pemodelan regresi linier yang baru tersebut, didapatkan nilai R_{new}^2
5. Sehingga statistik uji Breush-Pagan adalah

$$X_{hit}^2 = nR_{new}^2 \quad (16)$$

dimana

n adalah banyaknya data

- d. Menentukan titik kritis pengujian

Dengan α sebesar 0.05, H_0 ditolak jika $\chi_{hit}^2 > \chi_{\alpha,p}^2$ atau p-value < 0.05

- e. Menentukan kesimpulan hasil pengujian

III. Rangkuman

Berikut ini adalah rangkuman materi pada Bab IV:

- a) Residual merupakan deviasi atau penyimpangan antara data yang sebenarnya dengan data prediksi yang didapatkan model.
- b) Analisis residual artinya bahwa peneliti melakukan pengecekan terhadap asumsi yang harusnya terpenuhi.
- c) Secara umum asumsi yang harus dipenuhi oleh residual antara lain: residual berdistribusi normal, residual tidak mengalami autokorelasi, dan residual tidak mengalami heteroskedastisitas.
- d) Cara memeriksa asumsi residual dapat dilakukan dengan 2 cara yaitu secara visual melalui plot dan secara pengujian metode statistika.

- e) Penggunaan plot untuk memeriksa asumsi residual antara lain: 1) Penggunaan Plot Q-Q untuk memeriksa asumsi residual berdistribusi normal, 2) Penyusunan scatter plot residual dengan urutan pengamatan untuk memeriksa asumsi residual tidak terjadi autokorelasi, dan 3) Penyusunan scatter plot nilai prediksi dengan residual standardized untuk memeriksa asumsi residual tidak terjadi heteroskedastisitas.
- f) Penggunaan pengujian secara statistika untuk memeriksa asumsi residual antara lain: 1) Pengujian Kolmogorov-Smirnov untuk memeriksa asumsi residual berdistribusi normal, 2) Pengujian Durbin-Watson untuk memeriksa asumsi residual tidak terjadi autokorelasi, 3) Pengujian Breush-Pagan untuk memeriksa asumsi residual tidak terjadi heteroskedastisitas.

IV. Tutorial Metode

Berikut adalah tutorial metode materi bab IV yakni langkah-langkah menganalisis asumsi residual sehingga dapat menjawab tujuan pembelajaran yang ke-2 yaitu Dapat melakukan pengecekan apakah residual memenuhi asumsi regresi linier yang telah ditetapkan menggunakan bahasa pemrograman Python.

1. Studi Kasus Analisis Residual

Dataset yang digunakan untuk melakukan analisis residual ini adalah sama dengan data studi kasus analisis regresi linier sederhana yang dapat dilihat pada Tabel 5. Pada bab ini, kita ingin mengetahui apakah model regresi yang didapatkan telah memenuhi asumsi residualnya atau tidak.

2. Langkah-langkah Penyelesaian

Berikut adalah langkah-langkah penyelesaian studi kasus analisis residual menggunakan bahasa pemrograman Python.

A. Memeriksa Asumsi Residual Menggunakan Plot

Setelah mendapatkan residual pada model regresi dataset Tabel 5, untuk memvalidasi apakah benar bahwa model regresi yang kita

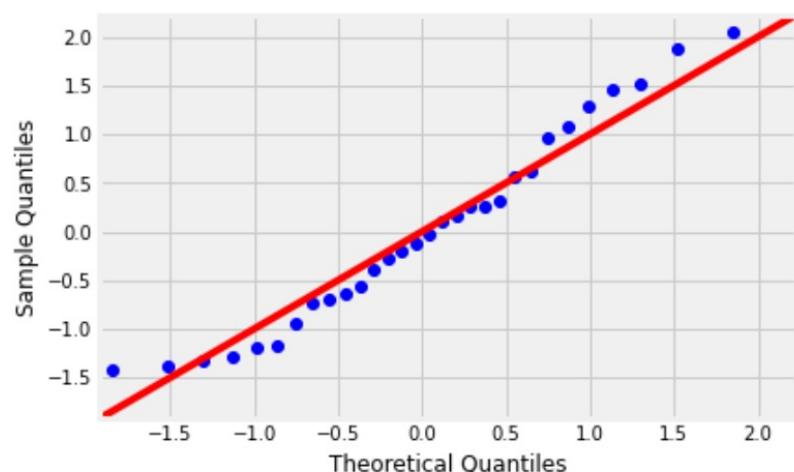
dapat sebelumnya dapat digunakan untuk memprediksi dan melihat pengaruh variabel lama bekerja terhadap pendapatan seseorang, maka harus dilakukan analisis residual. Langkah awal adalah memeriksa asumsi residual secara visual menggunakan suatu Plot. Berikut adalah kode pemrograman Python pada layanan Google Colab untuk memeriksa asumsi residual secara visual menggunakan suatu plot.

a. Memeriksa asumsi residual berdistribusi normal

Berikut adalah kode pemrograman Python menggunakan layanan Google Colab untuk memeriksa asumsi residual berdistribusi normal dengan Q-Q Plot. Kode dibawah merupakan kelanjutan dari kode pemrograman Python layanan Google Colab pada analisis regresi sederhana

```
#Menggambar Plot QQ menggunakan package statmodels
import scipy.stats as stats
fig = sm.qqplot(residual, stats.t, fit=True, line="45")
plt.show()
```

Setelah me-running kode pemrograman diatas, didapatkan hasil output sebagai berikut:



Gambar 15. Plot Q-Q Data Residual Pendapatan

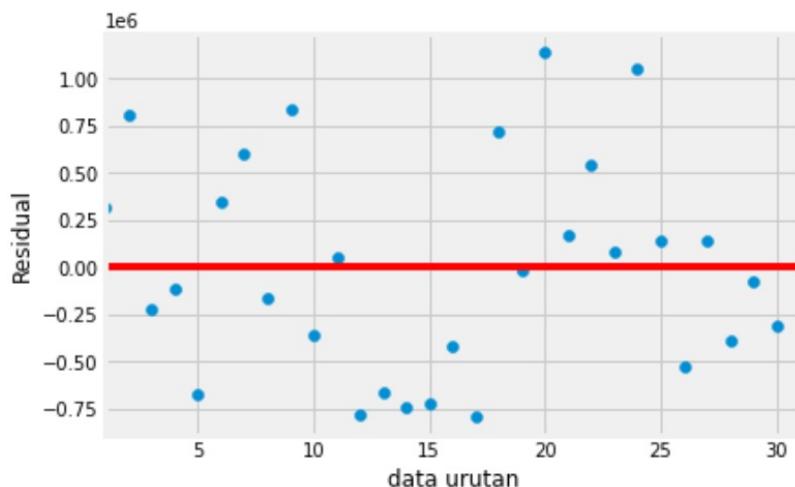
Berdasarkan Gambar 15 yang merupakan Plot Q-Q untuk memeriksa asumsi residual data Tabel 15 dapat dinyatakan bahwa titik-titik distribusi data tidak jauh dari garis linier, oleh karena itu dapat disimpulkan bahwa data residual pendapatan seseorang memenuhi asumsi berdistribusi normal

b. Memeriksa asumsi residual tidak terjadi autokorelasi

Berikut adalah kode pemrograman Python menggunakan layanan Google Colab untuk memeriksa asumsi residual tidak terjadi autokorelasi. Kode dibawah merupakan kelanjutan dari kode pemrograman Python layanan Google Colab pada analisis regresi sederhana setelah mendapatkan data residual.

```
#Plot memeriks asumsi residual tidak terjadi autokore
lasi
urutan_pengamatan=pd.Series(range(1,31))
plt.scatter(data_urutan, residual);
plt.axhline(0, color='red')
plt.xlabel('data urutan');
plt.ylabel('Residual');
plt.xlim([1,31]);
```

Setelah me-running kode pemrograman diatas, didapatkan hasil output sebagai berikut:



Gambar 16. Plot Residual dengan Urutan Pengamatan

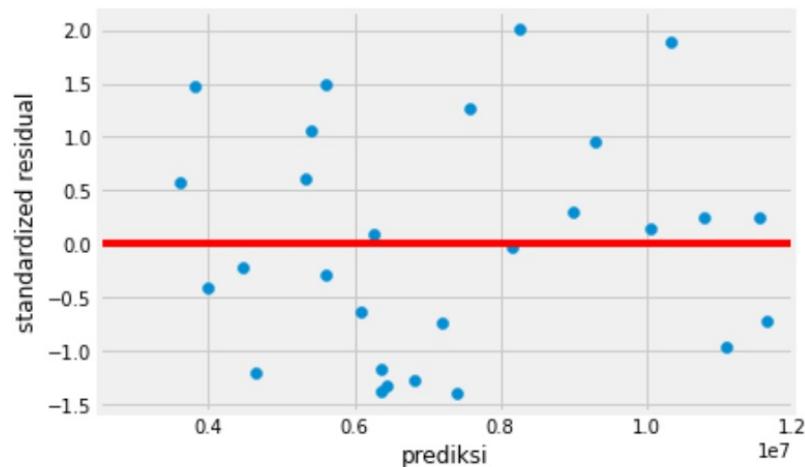
Berdasarkan Gambar 16 terlihat bahwa titik-titik data menyebar secara random dan tidak membentuk suatu pola, oleh karena itu dapat disimpulkan bahwa data residual pendapatan seseorang memenuhi asumsi tidak terjadi autokorelasi.

c. Memeriksa asumsi residual tidak terjadi heteroskedastisitas

Berikut adalah kode pemrograman Python menggunakan layanan Google Colab untuk memeriksa asumsi residual tidak terjadi heteroskedastisitas. Kode dibawah merupakan kelanjutan dari kode pemrograman Python layanan Google Colab pada analisis regresi sederhana setelah mendapatkan data residual.

```
#Plot memeriksa asumsi residual tidak terjadi heteroskedastisitas
influence = model.get_influence()
#menentukan standardized residualnya
std_residual = influence.resid_studentized_internal
plt.scatter(prediksi, std_residual);
plt.axhline(0, color='red')
plt.xlabel('prediksi');
plt.ylabel('standardized residual');
plt.xlim([2500000,12000000]);
```

Setelah me-running kode pemrograman diatas, didapatkan hasil output sebagai berikut:



Gambar 17. Plot Standardized Residual dengan Nilai Prediksi

Berdasarkan Gambar 17 terlihat bahwa titik-titik data menyebar secara random dan tidak membentuk suatu pola, oleh karena itu dapat disimpulkan bahwa data residual pendapatan seseorang memenuhi asumsi tidak terjadi heteroskedastisitas.

B. Memeriksa Asumsi Residual Menggunakan Metode Uji Statistika

Setelah melakukan pemeriksaan asumsi residual secara visual, langkah selanjutnya adalah melakukan pengujian secara statistika. Berikut adalah kode pemrograman Python pada layanan Google Colab untuk memeriksa asumsi residual dengan pengujian secara statistika.

c. Memeriksa asumsi residual berdistribusi normal

Berikut adalah kode pemrograman Python menggunakan layanan Google Colab untuk memeriksa asumsi residual berdistribusi normal dengan pengujian Jarque Bera. Kode dibawah merupakan kelanjutan dari kode pemrograman Python layanan Google Colab pada analisis regresi sederhana

```
#Pengujian Jarque
Bera untuk memeriksa asumsi residual
from statsmodels.compat import lzip
import statsmodels.formula.api as smf
import statsmodels.stats.api as sms
name = ["Jarque-Bera", "Chi^2 two-
tail prob.", "Skew", "Kurtosis"]
test = sms.jarque_bera(model.resid)
lzip(name, test)
```

Setelah me-running kode pemrograman diatas, didapatkan hasil output sebagai berikut.

```
[('Jarque-Bera', 1.5691038618767998),
('Chi^2 two-tail prob.', 0.4563241207058726),
('Skew', 0.3629411354039407),
('Kurtosis', 2.1465493462254344)]
```

Berdasarkan Hasil output diatas maka akan dilakukan langkah-langkah pengujian Jarque-Bera

- Menentukan hipotesis null dengan hipotesis alternatifnya

H_0 : Populasi berdistribusi normal

H_1 : Populasi tidak berdistribusi normal

- Menentukan level of significance atau α sebesar 5% atau 0.05
- Menentukan Statistik Uji

Berdasarkan output pemrograman python pengujian Jarque Bera didapatkan nilai statistik uji JB sebesar 1.5691 dan p-value sebesar 0.4563

- Menentukan titik kritis pengujian

Dengan α sebesar 0.05, H_0 ditolak jika $JB > \chi^2_{(\alpha,2)}$ dimana

$$\chi^2_{(\alpha,2)} = \chi^2_{(0.05,2)} = 5.99 \text{ atau } p\text{-value} < 0.05$$

- Menentukan kesimpulan hasil pengujian

Berdasarkan uji statistik didapatkan nilai JB sebesar 1.5691, dan p-value sebesar 0.4563. Dengan α sebesar 0.05 dan Tabel chi-square yaitu $\chi^2_{(\alpha,2)} = \chi^2_{(0.05,2)} = 5.99$, maka $JB < \chi^2_{(\alpha,2)}$ dan p-value > 0.05 yang dapat disimpulkan bahwa H_0 gagal ditolak artinya residual memenuhi asumsi berdistribusi normal.

- Memeriksa asumsi residual tidak terjadi autokorelasi

Berikut adalah kode pemrograman Python menggunakan layanan Google Colab untuk memeriksa asumsi residual tidak terjadi autokorelasi dengan pengujian Durbin-Watson. Kode dibawah merupakan kelanjutan dari kode pemrograman Python layanan Google Colab pada analisis regresi sederhana

```
#Pengujian Durbin-Watson
from statsmodels.stats.stattools import durbin_watson
dw = durbin_watson(model.resid)
print(f"Durbin-Watson: {dw}")
```

Setelah me-running kode pemrograman diatas, didapatkan hasil output yaitu

Durbin-Watson: 1.6479910076183397

- 1) Menentukan hipotesis null dengan hipotesis alternatifnya

H_0 : Tidak terjadi Autokorelasi

H_1 : Terjadi Autokorelasi

- 2) Menentukan level of significance atau α sebesar 5% atau 0.05

- 3) Menentukan Statistik Uji

Berdasarkan output diatas didapatkan bahwa statistic uji Durbin-Watsan adalah 1.648

- 4) Menentukan titik kritis pengujian

Dengan α sebesar 0.05 dan $n = 30$, maka kemungkinan kesimpulan pengujian ini adalah

H_0 ditolak jika $d_{hit} < d_L(1.3520)$,

H_0 gagal ditolak jika $d_{hit} > d_U(1.4688)$

Tidak dapat menarik kesimpulan $d_L \leq d \leq d_U$

- 5) Menentukan kesimpulan hasil pengujian

Dikarenakan d_{hit} sebesar 1.648 dimana lebih besar dari $d_U(1.4688)$ dapat disimpulkan bahwa H_0 gagal ditolak yang artinya bahwa asumsi residual tidak terjadi autokorelasi terpenuhi.

e. Memeriksa asumsi residual tidak terjadi heteroskedastisitas

Berikut adalah kode pemrograman Python menggunakan layanan Google Colab untuk memeriksa asumsi residual tidak terjadi autokorelasi dengan pengujian Breush-Pagan. Kode dibawah merupakan kelanjutan dari kode pemrograman Python layanan Google Colab pada analisis regresi sederhana

```
#Pengujian Breush-Pagan
from statsmodels.compat import lzip
import statsmodels.stats.api as sms
```

```

names = ['Lagrange multiplier statistic', 'p-value',
         'f-value', 'f p-value']
test = sms.het_breushpagan(model.resid, model.model.
                           exog)

lzip(names, test)

```

Setelah me-running kode pemrograman diatas, didapatkan hasil output yaitu

```

[('Lagrange multiplier statistic',
  0.3990532776649991),
 ('p-value', 0.5275785890839444),
 ('f-value', 0.3774707572507874),
 ('f p-value', 0.5439204509328637)]

```

- Menentukan hipotesis null dengan hipotesis alternatifnya
 H_0 : Tidak terjadi Heteroskedastisitas
 H_1 : Terjadi Heteroskedastisitas
- Menentukan level of significance atau α sebesar 5% atau 0.05
- Menentukan Statistik Uji

Berdasarkan output diatas didapatkan nilai Statistik Uji Breush-Pagan sebesar 0.3990 dan p-value sebesar 0.5276

- Menentukan titik kritis pengujian

Dengan α sebesar 0.05, H_0 ditolak jika $\chi^2_{hit} > \chi^2_{\alpha,p}$ atau p-value < 0.05 . $\chi^2_{0.05,1} = 3.841$

- Menentukan kesimpulan hasil pengujian

Dikarenakan Statistik Uji Breush-Pagan sebesar 0.3990 dimana lebih kecil dari $\chi^2_{0.05,1} = 3.841$ dapat disimpulkan bahwa H_0 gagal ditolak yang artinya bahwa asumsi residual tidak terjadi heteroskedastisitas.

V. Instruksi Tugas

Setelah memahami langkah-langkah analisis asumsi residual dengan bahasa pemrograman Python, pengguna e-modul praktikum dapat menyelesaikan tugas dengan instruksi sebagai berikut

- 1) Melakukan analisis data dengan bahasa pemrograman Python memanfaatkan layanan Google Colab
- 2) Hasil jawaban soal diketik menggunakan File Word ukuran A4 dengan format sebagai berikut:
 - i. Cover Laporan memuat Judul, Identitas pengguna, dan, Identitas Institusi
 - j. Daftar Isi, jika ada tabel dan gambar maka harus ada Daftar Tabel dan Daftar Gambar
 - k. Bab I Pendahuluan memuat latar belakang, tujuan, dan manfaat
 - l. Bab II Tinjauan Pustaka memuat teori dari metode
 - m. Bab III Metodologi data memuat dataset yang digunakan serta Langkah-langkah penyelesaian
 - n. Bab IV Hasil analisis memuat hasil penyelesaian berdasarkan soal permasalahan
 - o. Bab V Kesimpulan memuat hasil kesimpulan berdasarkan bab IV. Kesimpulan tidak boleh hasil dari copy paste
 - p. Bab VI Daftar Pustaka
- 3) Laporan dikumpulkan berbentuk Pdf.
- 4) File Notebook Google Colab hasil pemrograman Python juga ikut dikumpulkan dengan laporan
- 5) Laporan dan File Notebook Google Colab dimasukan kedalam satu folder dimana folder tersebut diberi nama identitas pengguna.

VI. Soal Asumsi Residual

Kerjakan analisis asumsi residual menggunakan data pada soal analisis regresi linier sederhana Tabel 6. Analisi asumsi residual ini dilakukan secara visual menggunakan plot dan menggunakan pengujian seperti pada sub bab tutorial metode Bab ini. Pengerjakan mengikuti Instruksi Tugas pada Sub Bab V.

BAB V

REGRESI LINIER BERGANDA

I. Tujuan Pembelajaran

Tujuan pembelajaran mahasiswa setelah selesai mempelajari bab V mengenai metode Regresi Linier Berganda antara lain:

- Dapat mengetahui teori metode regresi linier berganda
- Dapat mengetahui cara penyelesaian metode regresi linier sederhana dengan bahasa pemrograman Python
- Dapat melakukan analisis asumsi regresi linier berganda

II. Uraian Materi

Berikut adalah uraian materi bab V sehingga dapat menjawab tujuan pembelajaran yang telah ditetapkan:

1. Teori Regresi Linier Berganda

Regresi Linier Berganda merupakan metode regresi linier yang menyelidiki hubungan antara variabel dependen (atau respon) dan dua atau lebih variabel independen (atau prediktor). Sebelum melakukan analisis regresi linier berganda, dibuat scatter plot antara variabel dependen dan variabel independen untuk mengetahui apakah keduanya memiliki hubungan yang linier atau tidak. Berikut adalah model regresi linier berganda.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon \quad (16)$$

dimana

β_0 adalah intersep

$\beta_1, \beta_2, \dots, \beta_k$, adalah koefisien parameter variabel dependen sejumlah k

x_1, x_2, \dots, x_k adalah variabel independen sejumlah k

y adalah variabel dependen atau variabel respon

ε adalah error random yang memiliki asumsi $\varepsilon \sim iidn(0, \sigma^2)$ dimana iidn merupakan singkatan dari independen, identik, dan berdistribusi normal.

Berikut adalah langkah-langkah dalam menganalisis data menggunakan metode regresi linier berganda

- a) Membuat Scatter plot untuk mengetahui hubungan linier antara satu variabel dependen dengan satu variabel independennya sejumlah variabel independennya.
- b) Mengetahui seberapa kuat hubungan korelasi antara variabel dependen dengan variabel dependennya menggunakan metode korelasi Pearson sejumlah variabel independennya,
- c) Melakukan estimasi untuk mendapatkan nilai estimator $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$
- d) Melakukan pengujian hipotesis Uji F untuk melihat apakah secara bersama-sama variabel independen yang lebih dari satu berpengaruh terhadap variabel dependennya,
- e) Melakukan perhitungan R-squared untuk melihat seberapa besar kesesuaian model yang didapatkan
- f) Melakukan pengujian hipotesis Uji t untuk melihat apakah secara parsial variabel independen berpengaruh terhadap variabel dependennya. Sehingga uji t ada sebanyak variabel independennya.
- g) Membuat kesimpulan dan interpretasi model
- h) Melakukan analisis asumsi residual dan asumsi multikolinieritas.
- i) Menghitung nilai prediksi dari model yang didapatkan

2. Teori Regresi Linier Sederhana

Berikut ini adalah penjelasan mengenai teori regresi linier berganda yang lebih detail.

A. Hubungan Linier Variabel Independen dengan Dependen

Asumsi yang paling dasar dalam analisis regresi adalah adanya hubungan yang linier antara variabel independen dengan variabel dependennya. Untuk melihat hubungan yang linier antara kedua variabel dapat menggunakan *scatter plot*. Pada regresi linier berganda banyaknya scatter plot yang dihasilkan tergantung dengan sejumlah variabel independen yang diketahui. Misal, ada 3 variabel independen

maka Analisa hubungan linier akan ada 3 scatter plot antara variabel dependen dengan masing-masing variabel dependennya.

B. Estimasi Model Regresi Linier Berganda

Diketahui bahwa model regresi linier sederhana dengan pengamatan $i = 1, 2, \dots, n$ adalah sebagai berikut.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i \quad (17)$$

dimana

β_0 adalah intersep

$\beta_1, \beta_2, \dots, \beta_k$, adalah koefisien parameter variabel dependen sejumlah k

$x_{i1}, x_{i2}, \dots, x_{ik}$ adalah variabel independen sejumlah k

y_i adalah variabel dependen atau variabel respon

ε_i adalah error random yang memiliki asumsi $\varepsilon_i \sim iidn(0, \sigma^2)$

Untuk mendapatkan estimator $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ sebaiknya dilakukan dengan mengubah persamaan (17) kedalam bentuk matriks untuk memudahkan perhitungan estimasi. Hasilnya adalah sebagai berikut.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (18)$$

Dimana

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Kriteria metode OLS menggunakan pendekatan matriks adalah

$$\min(S(\boldsymbol{\beta})) = \min\left(\sum_{i=1}^n \varepsilon_i^2\right) = \min(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}) = \min\left((\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right) \quad (19)$$

Berdasarkan kriteria metode estimasi OLS diatas dengan menggunakan perhitungan penuruan persamaan (19) yang akan disama dengan nol (nol) maka hasil dari estimator $\hat{\boldsymbol{\beta}}$ adalah

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} \quad (20)$$

Setelah mendapatkan estimator $\hat{\boldsymbol{\beta}}$ maka model regresi hasil estimasi yang didapat yaitu

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \dots + \hat{\beta}_k x_{ki} \quad (21)$$

$$\text{atau } \hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

dimana \mathbf{H} matriks $n \times n$ yaitu $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ disebut matriks Hat.

Oleh karena itu didapat residual yang merupakan selisih antara nilai aktual dengan nilai prediksi berdasarkan variabel dependennya. Berikut ini adalah persamaan residual yang didapat,

$$\begin{aligned}\mathbf{e} &= \mathbf{y} - \hat{\mathbf{y}} \\ \mathbf{e} &= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}\end{aligned}\tag{22}$$

C. Pengujian Hipotesis Uji F

Uji F pada analisis regresi linier berganda digunakan untuk mengetahui apakah secara bersama-sama ada minimal satu variabel independen yang berpengaruh terhadap variabel dependennya. Berikut adalah Langkah-langkah dalam melakukan Uji Hipotesis pada Uji F

- Menentukan hipotesis null dengan hipotesis alternatifnya

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1: \text{minimal ada satu } j \text{ yang } \beta_j \neq 0 \text{ dimana } j = 1, 2, \dots, k$$

- Menentukan level of significance atau α sebesar 5% atau 0.05

- Menentukan Uji Statistik F

Tabel 7. Uji Statistik F

	SS	df	MS	F
Regresi	SS_{Reg}	k	$MS_{Reg} = \frac{SS_{Reg}}{k}$	$F_{hit} = \frac{MS_{reg}}{MS_{res}}$
Residual	SS_{Res}	$n - k - 1$	$MS_{Res} = \frac{SS_{Res}}{n - k - 1}$	
Total	SS_{Tot}	$n - 1$		

- Menentukan titik kritis pengujian

Dengan α sebesar 0,05, H_0 ditolak jika $F_{hit} > F_{tabel}$ dimana $F_{tabel} = F_{\alpha, 1, n-p}$ dimana $p = 1 + k$ atau p-value < 0,05.

- Menentukan kesimpulan hasil pengujian

D. Nilai R-Squared

Perhitungan R-squared digunakan untuk melihat seberapa besar kesesuaian model yang didapatkan. Perhitungan ini akan melihat

seberapa besar pengaruh variabel independen terhadap dapat variabel dependennya. Berikut adalah rumus dari R-Squared

$$R^2 = \frac{SS_{\text{Reg}}}{SS_{\text{Tot}}} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{SS_{\text{Res}}}{SS_{\text{Tot}}} \quad (23)$$

Nilai R Squared akan terus bertambah mengikuti banyaknya variabel independen. Oleh karenanya, ketika terdapat lebih dari satu variabel independen, lebih disarankan menggunakan Adjusted R^2 karena perhitungan ini sudah mempertimbangkan banyaknya variabel

Nilai Adjusted R-Squared adalah antara 0 sampai dengan 1 yang merupakan proporsi dari total variansi yang dijelaskan oleh model regresi.

E. Pengujian Hipotesis Uji t

Uji t pada analisis regresi linier berganda digunakan untuk mengetahui apakah masing-masing variabel independen berpengaruh terhadap variabel dependennya. Pengujian ini dikatakan sebagai pengujian signifikansi parameter secara parsial. Berikut adalah Langkah-langkah dalam melakukan Uji Hipotesis pada Uji t

1. Menentukan hipotesis null dengan hipotesis alternatifnya

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

dimana $j = 1, 2, \dots, k$

2. Menentukan level of significance atau α sebesar 5% atau 0.05

3. Menentukan Uji Statistik t-hitung

$$t_{\text{hit}} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \quad (24)$$

dimana $j = 1, 2, \dots, k$

dimana

$se(\hat{\beta}_j)$ adalah estimasi standar error

4. Menentukan titik kritis pengujian

Dengan α sebesar 0.05, H_0 ditolak jika $|t_{hit}| > t_{tabel}$ dimana $t_{tabel} = t_{\frac{\alpha}{2}, n-p}$ diman $p = 1 + k$ atau p-value < 0.05

5. Menentukan kesimpulan hasil pengujian

F. Pengujian Asumsi Regresi

Asumsi yang harus dipenuhi dalam analisis regresi linier berganda antara lain: a) Adanya hubungan yang linier antara variabel dependen dengan masing-masing variabel independennya, b) Residual yang didapat berdistribusi normal, c) Residual yang didapat tidak terjadi autokorelasi, d) Residual yang didapat tidak terjadi heteroskedastisitas, dan e) tidak terjadi multikolinieritas pada variabel independennya. Penjelasan asumsi b) sampai d) dapat dilihat pada Bab III mengenai Asumsi Residual. Adapun pada Bab ini akan dijelaskan mengenai asumsi variabel independen tidak boleh terjadi multikolinieritas.

Multikolinieritas merupakan suatu penyimpangan dimana telah terjadi hubungan yang kuat atau berkorelasi antar variabel independennya. Ketika antar variabel berkorelasi maka akan menyebabkan dua masalah yaitu koefisien menjadi sangat sensitif terhadap perubahan kecil dalam model dan dapat mengurangi ketepatan estimasi koefisien.

Pengujian multikolinieritas biasanya menggunakan nilai dari Variance Inflation Factor atau biasanya disingkat nilai VIF. Rumus dari VIF adalah sebagai berikut.

$$VIF_j = \frac{1}{1 - R_j^2} \quad (25)$$

Dimana R_j^2 adalah nilai R^2 yang diperoleh dengan meregresikan prediktor ke- j pada prediktor yang tersisa.

Suatu variabel independen dapat dikatakan secara signifikan terjadi multikolinieritas adalah ketika memiliki nilai VIF lebih dari 10.

III. Rangkuman

Berikut ini adalah rangkuman materi pada Bab II:

- a) Regresi Linier Berganda merupakan metode regresi linier yang menyelidiki hubungan antara variabel dependen (atau respon) dan dua atau lebih variabel independen (atau prediktor).
- b) Berikut adalah langkah-langkah analisis regresi berganda antara lain: membuat scatter plot variabel dependen dengan independen, menghitung nilai korelasi variabel dependen dengan independen, mendapatkan estimator $\hat{\beta}$, melakukan Uji F, mendapatkan R-Squared, dan melakukan Uji T, memeriksa asumsi regresi linier berganda, membuat kesimpulan dan interpretasi model, dan menghitung nilai prediksi dari model yang didapatkan.
- c) Pengujian asumsi pada regresi linier berganda dimana tidak boleh terjadi multikolinieritas dapat diperiksa melalui nilai VIF. Ketika nilai $VIF > 10$ maka dapat disimpulkan telah terjadi multikolinieritas.

IV. Tutorial Metode

Berikut adalah tutorial metode materi bab V yakni langkah-langkah analisis regresi kinier berganda sehingga dapat menjawab tujuan pembelajaran yang ke-2 dan ke-3 yaitu dapat mengetahui cara penyelesaian metode regresi linier berganda serta dapat melakukan analisis asumsinya dengan bahasa pemrograman Python.

1. Studi Kasus Regresi Linier Berganda

Suatu program studi mepunyai data lama lulus mahasiswa Angkatan yang pertama. Terdapat penelitian terdahulu yang mengatakan bahwa lama lulus mahasiswa dipengaruhi oleh IQ dan nilai IPK. Oleh karena itu, Prodi di suatu universitas ingin mengetahui apakah benar bahwa lama lulus mahasiswa dipengaruhi oleh kedua variabel tersebut. Ketika benar-benar dipengaruhi, progdi ingin memprediksi lama lulus mahasiswa jika

memiliki IQ sebesar 110 dan IPK semester 5 sebesar 3.25. Penelitian akan dilakukan dengan menggunakan analisis regresi berganda sampai dengan memeriksa asumsinya. Berikut adalah data yang akan digunakan.

Tabel 8. Data Lama Lulus dan Faktor yang Mempengaruhi

Data ke-1	Lama Lulus (Tahun)	IPK	IQ
1	5	3.12	108
2	3.5	3.65	129
3	3.5	3.52	130
4	3.5	3.82	135
5	4.5	3.25	110
6	5	3.11	105
7	5	3.15	106
8	4	3.53	115
9	4	3.55	116
10	4	3.57	118
11	3.5	3.76	130
12	3.5	3.75	130
13	4.5	3.22	112
14	4	3.27	113
15	4	3.25	113
16	3.5	3.82	133
17	3.5	3.86	134
18	3.5	3.9	131
19	3.5	4	132
20	5	3.02	108
21	5	3.29	108
22	4.5	3.26	111
23	4.5	3.27	112
24	3.5	4	134
25	4	3.75	118
26	4	3.55	113
27	4	3.54	113
28	3.5	3.75	122
29	3.5	3.75	122
30	5	2.95	108

2. Langkah-langkah Penyelesaian

Berikut adalah-langkah-langkah analisis regresi linier berganda untuk menyelesaikan permasalahan pada studi kasus data Tabel 8 menggunakan bahasa pemrograman Python.

A. Menentukan Variabel Independen dan Variabel Dependend

Berdasarkan sub bab studi kasus, diketahui bahwa Program Studi di suatu universitas ingin apakah benar bahwa besarnya IQ dan nilai IPK dapat mempengaruhi lama lulus mahasiswa. Dan juga diketahui bahwa Program Studi ingin memprediksi berapa mahasiswa lulus jika IQ yang dimiliki sebesar 110 dan IPK semester 5 sebesar 3.25. Dari permasalahan ini, maka dapat diketahui bahwa terdapat dua variabel independen yaitu IQ (x_1) dan IPK (x_2) dan satu variabel dependen yaitu Lama Lulus Mahasiswa (y)

B. Memasukkan Data ke Notebook Google Colab

Dikarenakan jumlah data yang lebih dari 20, tidak disarankan untuk mengetik secara langsung pada Notebook Google Colab. Hal ini karena akan memakan waktu yang lama dan tidak efisien. Oleh karena itu data Tabel 8 disimpan terlebih dahulu ke bentuk File csv yang Bernama : **lama_lulus**

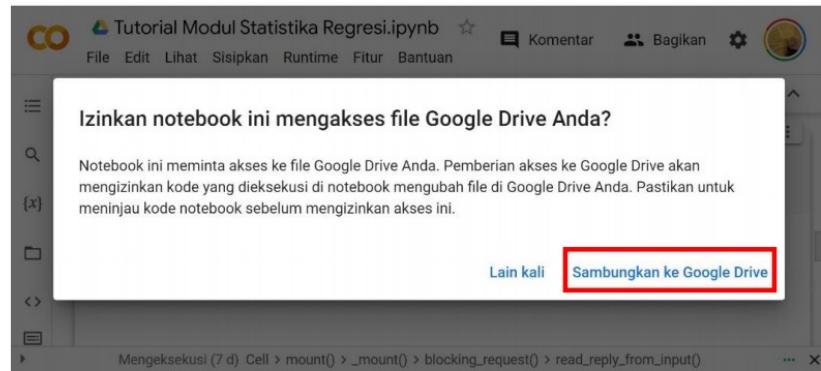
Pada e-modul Praktikum Statistika Regresi ini, dalam memanggil data bentuk File csv terlebih dahulu disimpan di Folder Google Drive masing-masing pengguna e-modul. Penulis menyimpan File csv pada Folder yang bernama : **Dataset eModul**.

Berikut adalah kode pemrograman python pada notebook Google Colab untuk memasukan data ke Notebook Google Colab

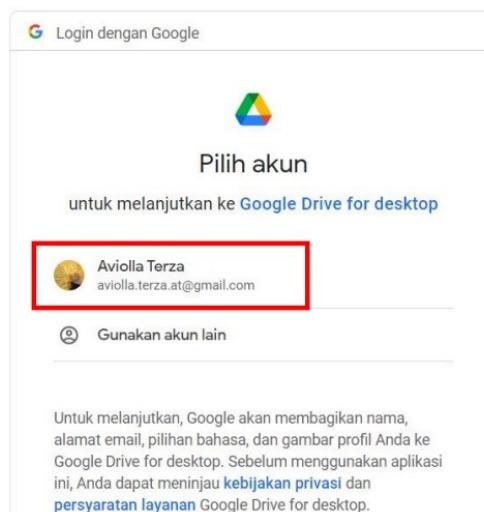
3) Menghubungkan Google Drive dengan Notebook Google Colab

```
#CONNECT GOOGLE DRIVE
from google.colab import drive
drive.mount('/content/drive')
```

Setelah me-running kode diatas akan muncul Gambar yang dapat dilihat pada Gambar 18 dan klik **Sambungkan ke Google Drive**. Kemudian akan muncul Gambar 19 untuk memilih akun Google Drive tempat meyimpan dataset kita: **Klik Akun → Izinkan**. Setelah itu dataset di Google Drive sudah dapat terbaca oleh Notebook Google Colab yang ditandai dengan kode : Mounted at /content/drive



Gambar 18. Izin Mengakses Google Drive Data Regresi Berganda



Gambar 19. Memilih Akun Google Drive untuk Data Regresi Berganda

4) Membuat dataframe data menggunakan library Pandas

Setelah data terhubung dari Google Drive ke Notebook Google Colab, Selanjutnya adalah menyusun dataset kita menjadi data frame dengan library Pandas. Kode pemrograman python-nya adalah

```
# Memanggil dataset
import pandas as pd
df = pd.read_csv("drive/MyDrive/Dataset eModul/lama_1
ulus.csv")
df.head()
```

Kode diatas akan menghasilkan Output yaitu

	IQ	IPK	Tahun Lulus di Universitas	
0	108	3.12	5.0	
1	129	3.65	3.5	
2	130	3.52	3.5	
3	135	3.82	3.5	
4	110	3.25	4.5	

Gambar 20. Output Data Lama Lulus Mahasiswa

Diketahui bahwa untuk mengefisienkan penamaan kode pemrograman python, maka nama variabel Gambar 20 diganti dari **IQ** menjadi x_1 dan **IPK** menjadi x_2 , serta **Tahun Lulus di Universitas** diganti menjadi y . Berikut adalah kode pemrogramannya

```
#Mengganti nama variabel
df.rename(columns={'IQ':'x1','IPK':'x2','Tahun Lulus
di Universitas':'y'}, inplace=True)
df.head()
```

Kode diatas akan menghasilkan output yaitu

	x1	x2	y	
0	108	3.12	5.0	
1	129	3.65	3.5	
2	130	3.52	3.5	
3	135	3.82	3.5	
4	110	3.25	4.5	

Gambar 21. Mengganti Nama Variabel di Data Lama Lulusan

C. Mengetahui Hubungan Linier Variabel Dependen dan Independen

Untuk mengetahui apakah variabel x dan variabel y mempunyai hubungan yang linier atau tidak dapat dilihat dari Plot *Scatter Plot*. Dikarenakan terdapat dua variabel independen, maka akan ada dua scatter plot masing-masing variabel dependen dan indpeenndenya.

Kode pemrograman python untuk membuat plot *Scatter Plot* masing-masing variabel independen adalah sebagai berikut.

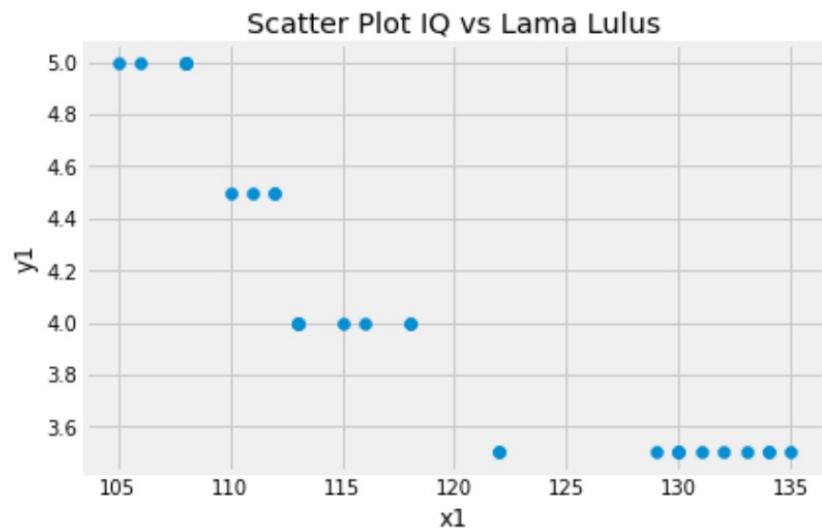
```
# Library untuk memunculkan Plot
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
plt.style.use('fivethirtyeight')
import warnings
warnings.filterwarnings('ignore')
%matplotlib inline

# Untuk memunculkan Scatter Plot IQ VS Lama Lulus
plt.scatter(df['x1'], df['y'])
plt.xlabel('x1')
plt.ylabel('y1')
plt.title('Scatter Plot IQ vs Lama Lulus')
plt.show()

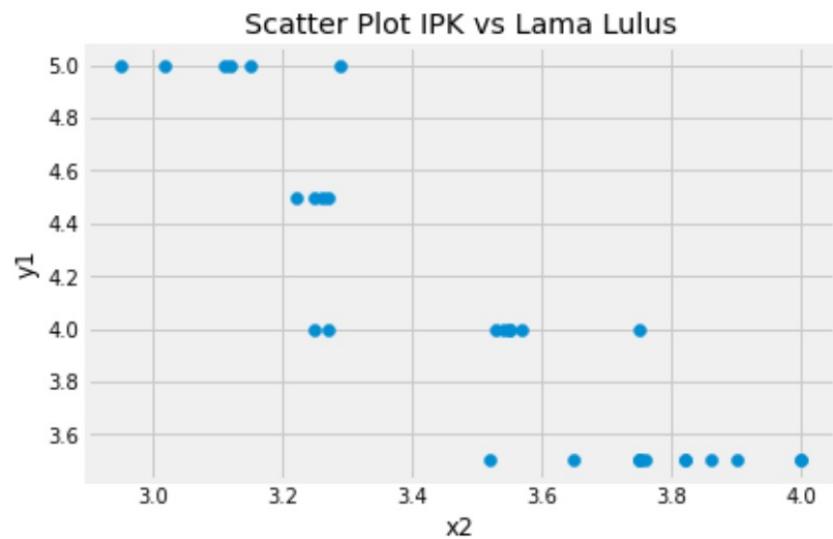
# Untuk memunculkan Scatter Plot IPK VS Lama Lulus
plt.scatter(df['x2'], df['y'])
plt.xlabel('x2')
plt.ylabel('y1')
plt.title('Scatter Plot IPK vs Lama Lulus')
plt.show()
```

Output dari Plot *Scatter plot* masing-masing variabel independen dengan variabel dependennya dapat dilihat pada Gambar 22 dan Gambar 23.

Berdasarkan Gambar 22 dapat dilihat bahwa terdapat hubungan linier yang negative antara Data IQ dengan Lama Lulus. Semakin besar IQ yang dimiliki mahasiswa maka semakin cepat mahasiswa tersebut dapat lulus dari Program Studi tersebut. Sehingga nantinya estimasi koefisien parameter pada regresi linier berganda diduga memiliki tanda yang negatif.



Gambar 22. Scatter Plot Data IQ vs Lama Lulus



Gambar 23. Scatter Plot Data IPK vs Lama Lulus

Berdasarkan Gambar 23 dapat dilihat bahwa terdapat hubungan linier yang negatif antara Data IPK dengan Lama Lulus. Semakin besar IPK yang didapatkan mahasiswa maka semakin cepat mahasiswa tersebut dapat lulus dari Program Studi tersebut. Sehingga nantinya estimasi koefisien parameter pada regresi linier berganda diduga memiliki tanda yang negatif.

D. Mengetahui Seberapa Kuat Hubungan Variabel Dependen dan Independen Menggunakan Metode Korelasi

Setelah mengetahui bahwa variabel x_1 dan x_2 mempunyai hubungan yang linier terhadap variabel y , langkah selanjutnya adalah mengetahui seberapa kuat hubungannya dengan menggunakan metode korelasi Pearson. Berikut adalah kode pemrograman python yang digunakan untuk mengetahui seberapa kuat hubungan variabel IQ (x_1) terhadap variabel lama lulus (y).

```
#Mengitung nilai korelasi pearson
from scipy.stats import pearsonr
# Convert dataframe into series
list1 = df['x1']
list2 = df['y']
corr, _ = pearsonr(list1, list2)
print('Koefisien Pearson: %.5f' % corr)
```

Hasil output dari kode diatas adalah

```
Koefisien Pearson: -0.89798
```

Berdasarkan nilai korelasi Pearson sebesar -0.89798 maka dapat disimpulkan bahwa variabel IQ dengan variabel lama lulus memiliki hubungan korelasi negatif yang kuat karena memiliki nilai koefisien korelasi lebih dari -0.8 .

Berikut adalah kode pemrograman python yang digunakan untuk mengetahui seberapa kuat hubungan variabel IPK (x_2) terhadap variabel lama lulus (y).

```
#Mengitung nilai korelasi pearson
from scipy.stats import pearsonr
# Convert dataframe into series
list1 = df['x2']
list2 = df['y']
corr, _ = pearsonr(list1, list2)
print('Koefisien Pearson: %.5f' % corr)
```

Hasil output dari kode diatas adalah

```
Koefisien Pearson: -0.90250
```

Berdasarkan nilai korelasi Pearson sebesar -0.90250 maka dapat disimpulkan bahwa variabel IPK dengan variabel lama lulus memiliki hubungan korelasi negatif yang kuat karena memiliki nilai koefisien korelasi lebih dari -0.8 .

E. Membentuk Model Regresi Linier Berganda

Langkah selanjutnya memodelkan dataset dengan menggunakan analisis regresi linier berganda. Kode pemrograman python untuk memodelkan data dengan regresi linier berganda adalah sebagai berikut.

```
#Memodelkan dengan Regresi Linier Berganda
import numpy as np
import statsmodels.api as sm
x = df[['x1','x2']]
y = df['y']
x = sm.add_constant(x)
model = sm.OLS(y, x).fit()
print_model = model.summary()
print(print_model)
```

Hasil output dari Kode pemrograman Python diatas adalah

```
OLS Regression Results
=====
Dep. Variable:          y    R-squared:       0.856
Model:                 OLS   Adj. R-squared:   0.846
Method:                Least Squares   F-statistic:     80.41
Date:      Thu, 17 Nov 2022   Prob (F-statistic):  4.24e-12
Time:          05:16:22   Log-Likelihood:   -3.2075
No. Observations:      30   AIC:             -0.4149
Df Residuals:          27   BIC:             3.789
Df Model:                  2
Covariance Type:    nonrobust
=====
            coef    std err        t     P>|t|      [0.025      0.975]
-----  
const    10.5132    0.511    20.569      0.000     9.464    11.562
x1      -0.0262    0.009    -2.800      0.009    -0.045    -0.007
x2      -0.9480    0.310    -3.061      0.005    -1.583    -0.313
=====
Omnibus:           0.507   Durbin-Watson:   1.262
Prob(Omnibus):    0.776   Jarque-Bera (JB):  0.614
Skew:              -0.253   Prob(JB):       0.736
Kurtosis:          2.516   Cond. No.    1.47e+03
=====

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.47e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Gambar 24. Output Model Regresi Linier Berganda

Berdasarkan Gambar 24 didapatkan model regresi linier berganda adalah

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \quad (26)$$

$$\hat{y} = 10.51 - 0.03x_1 - 0.95x_2$$

Dari Persamaan (13), kita dapat menghasilkan nilai prediksi dan residualnya. Kode pemrograman python untuk menghasilkan nilai prediksi \hat{y} adalah

```
prediksi = model.predict(x)
print(prediksi.head())
```

Output hasil kode pemrograman diatas antara lain

```
0    4.721905
1    3.668495
2    3.765500
3    3.349915
4    4.546190
dtype: float64
```

Adapun kode pemrograman python untuk menghasilkan nilai residual e_i adalah

```
residual=model.resid
print(residual.head())
```

Output hasil kode pemrograman diatas antara lain

```
0    0.278095
1   -0.168495
2   -0.265500
3    0.150085
4   -0.046190
dtype: float64
```

F. Melakukan Uji F

Setelah mendapatkan model regresi berganda, langkah selanjutnya adalah mengetahui apakah secara bersama-sama ada minimal satu variabel independen yang berpengaruh terhadap variabel

dependennya. Untuk itu diperlukan suatu Pengujian Hipotesis Uji F dengan langkah-langkah sebagai berikut.

1. Menentukan hipotesis null dengan hipotesis alternatifnya

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

$$H_1 : \text{minimal ada satu } \beta_j \neq 0, \text{ dimana } j = 1, 2$$

2. Menentukan level of significance atau α sebesar 5% atau 0.05

3. Menentukan Uji Statistik F

Pada Gambar 24, didapatkan bahwa Nilai F-statistic yang diperoleh adalah 80.41 atau p-value yang didapat sebesar 4.24e-12

4. Menentukan titik kritis pengujian

Dengan α sebesar 0,05, H_0 ditolak jika $F_{hit} > F_{tabel}$ dimana $F_{tabel} = F_{0.05,1,27} =$ atau p-value < 0,05.

5. Menentukan kesimpulan hasil pengujian

Dengan α sebesar 0,05, nilai $F_{hit} = 622.5$ lebih besar dibandingkan $F_{tabel} = F_{0.05,1,28} = 3.3541$ dan juga nilai p-value yang didapatkan sebesar 0.00 yang kurang dari 0,05. Berdasarkan hal itu maka dapat disimpulkan bahwa H_0 ditolak yang artinya minimal ada satu variabel independen yang berpengaruh terhadap variabel dependennya.

G. Mendapatkan Nilai R-Squared

Langkah selanjutnya adalah mengetahui nilai proporsi dari total varians yang dijelaskan oleh model regresi atau dikenal sebagai nilai R-Squared. Pada analisis regresi linier berganda, karena memiliki dua variabel independen maka lebih baik menggunakan R-Squared Adjusted untuk melihat seberapa besar variabel dependen dapat dijelasakan oleh variabel independennya. Berdasarkan Gambar 24, didapatkan nilai R-Squared Adjusted sebesar 0.846 yang artinya bahwa 84,6 % lama lulus mahasiswa dipengaruhi oleh IQ dan IPK mahasiswa dan sisanya 15,4% dipengaruhi oleh variabel lainnya yang tidak diketahui atau tidak diteliti.

H. Melakukan Uji T

Langkah selanjutnya adalah melihat apakah variabel independen berpengaruh secara signifikan atau tidak terhadap variabel dependennya. Dikarenakan variabel independen berjumlah dua, maka Uji T akan dilakukan sebanyak dua kali. Langkah-langkah Uji T untuk variabel IQ adalah sebagai berikut.

1. Menentukan hipotesis null dengan hipotesis alternatifnya

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

2. Menentukan level of significance atau α sebesar 5% atau 0.05

3. Menentukan Uji Statistik t-hitung

Pada Gambar 24, didapatkan bahwa Nilai t -statistic untuk variabel IQ adalah -2.800 atau p-value yang didapat sebesar 0.009

4. Menentukan titik kritis pengujian

Dengan α sebesar 0.05, H_0 ditolak jika $|t_{hit}| > t_{tabel}$ dimana $t_{tabel} = t_{\frac{\alpha}{2}, n-3} = t_{0.025; 27} = 2.052$

5. Menentukan kesimpulan hasil pengujian

Dengan α sebesar 0,05, nilai $|t_{hit}| = 2.800$ lebih besar dibandingkan $t_{tabel} = 2.052$ dan juga nilai p-value yang didapatkan sebesar 0.009 yang kurang dari 0,05. Berdasarkan hal itu maka dapat disimpulkan bahwa H_0 ditolak yang artinya bahwa variabel IQ mahasiswa berpengaruh signifikan terhadap lama lulusnya.

Berikut adalah langkah-langkah Uji T untuk variabel IPK apakah variabel tersebut berpengaruh signifikan terhadap variabel lama lulus.

1. Menentukan hipotesis null dengan hipotesis alternatifnya

$$H_0 : \beta_2 = 0$$

$$H_1 : \beta_2 \neq 0$$

2. Menentukan level of significance atau α sebesar 5% atau 0.05

3. Menentukan Uji Statistik t-hitung

Pada Gambar 24, didapatkan bahwa Nilai t -statistic untuk variabel IQ adalah -3.061 atau p-value yang didapat sebesar 0.005

4. Menentukan titik kritis pengujian

Dengan α sebesar 0.05 , H_0 ditolak jika $|t_{hit}| > t_{tabel}$ dimana $t_{tabel} = t_{\frac{\alpha}{2}, n-3} = t_{0.025; 27} = 2.052$

5. Menentukan kesimpulan hasil pengujian

Dengan α sebesar 0.05 , nilai $|t_{hit}| = 3.061$ lebih besar dibandingkan $t_{tabel} = 2.052$ dan juga nilai p-value yang didapatkan sebesar 0.005 yang kurang dari 0.05 . Berdasarkan hal itu maka dapat disimpulkan bahwa H_0 ditolak yang artinya bahwa variabel IPK mahasiswa berpengaruh signifikan terhadap lama lulusnya.

I. Uji Asumsi Regresi

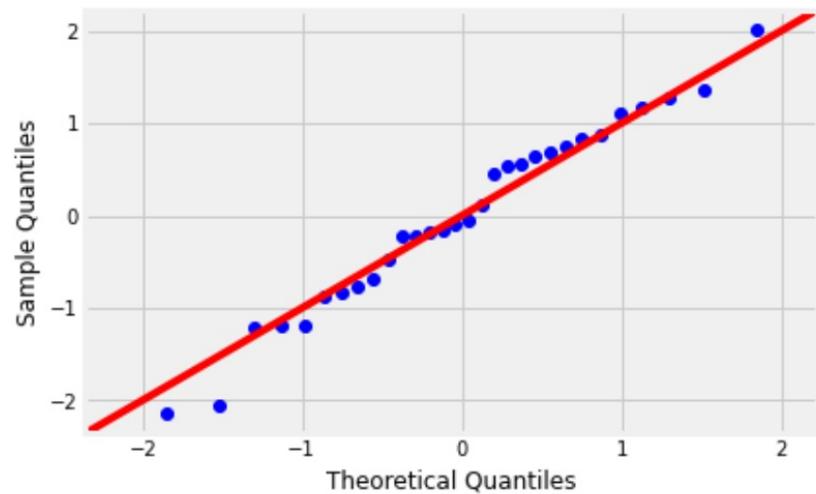
Setelah mendapatkan residual pada model regresi linier berganda pada data lama lulus mahasiswa, langkah selanjutnya adalah memeriksa asumsi residual dan memeriksa asumsi tidak terjadinya multikolinieritas. Berikut adalah langkah-langkah menganalisis asumsi residual menggunakan bahasa pemrograman Python pada layanan Google Colab menggunakan suatu plot

a. Memeriksa asumsi residual berdistribusi normal

Berikut adalah kode pemrograman Python menggunakan layanan Google Colab untuk memeriksa asumsi residual berdistribusi normal dengan Q-Q Plot.

```
#Menggambar Plot QQ menggunakan package statmodels
import scipy.stats as stats
fig = sm.qqplot(residual, stats.t, fit=True, line="45")
plt.show()
```

Setelah me-running kode pemrograman diatas, didapatkan hasil output sebagai berikut:



Gambar 25. Plot Q-Q data Residual Lama Lulusan

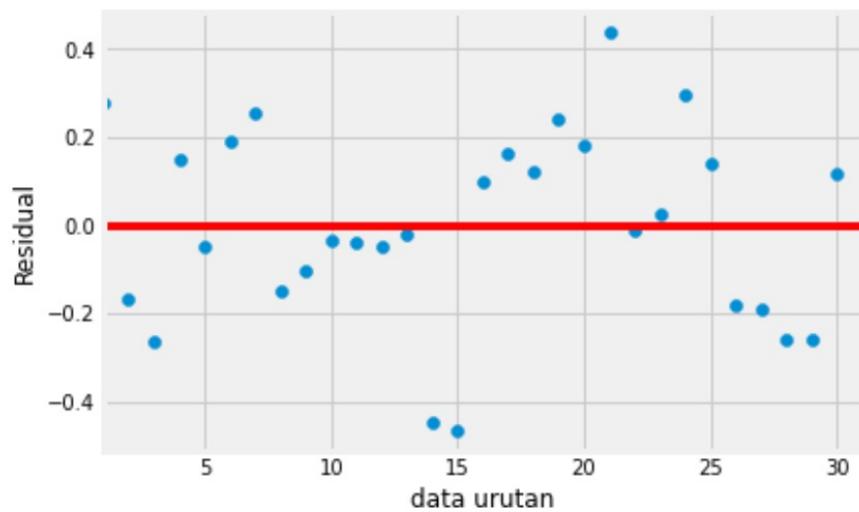
Gambar 25 terlihat bahwa distribusi data residual mendekati garis linier. Sehingga dapat disimpulkan bahwa data residual tersebut memenuhi asumsi berdistribusi normal.

b. Memeriksa asumsi residual tidak terjadi autokorelasi

Berikut adalah kode pemrograman Python menggunakan layanan Google Colab untuk memeriksa asumsi residual tidak terjadi autokorelasi.

```
#Plot memeriks asumsi residual tidak terjadi autokorelasi
urutan_pengamatan=pd.Series(range(1,31))
plt.scatter(urutan_pengamatan, residual);
plt.axhline(0, color='red')
plt.xlabel('data urutan');
plt.ylabel('Residual');
plt.xlim([1,31]);
```

Setelah me-running kode pemrograman diatas, didapatkan hasil output sebagai berikut:



Gambar 26. Plot Residual Lama Lulusan dengan Data urutannya

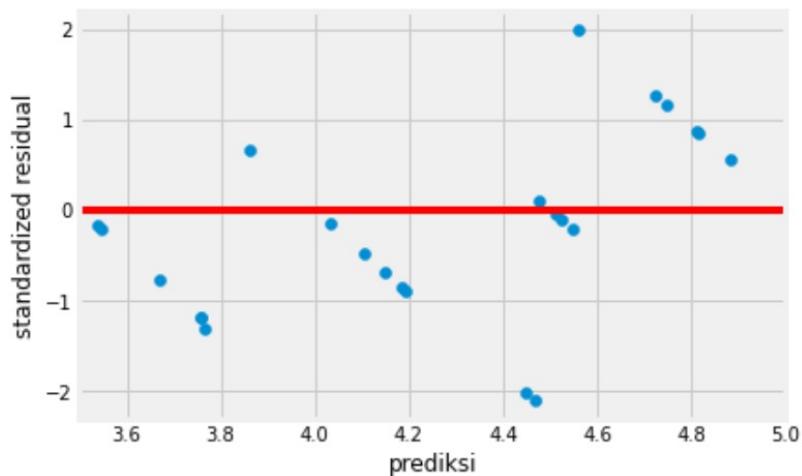
Gambar 26 terlihat bahwa distribusi data residual terhadap plot membentuk suatu pola yang naik dan menurun, Sehingga dapat disimpulkan bahwa data residual telah terjadi autokorelasi yang artinya residual ini tidak memenuhi asumsi.

- c. Memeriksa asumsi residual tidak terjadi heteroskedastisitas

Berikut adalah kode pemrograman Python menggunakan layanan Google Colab untuk memeriksa asumsi residual tidak terjadi heteroskedastisitas.

```
#Plot memeriksa asumsi residual tidak terjadi heteroskedastisitas
influence = model.get_influence()
#menentukan standardized residualnya
std_residual = influence.resid_studentized_internal
plt.scatter(prediksi, std_residual);
plt.axhline(0, color='red')
plt.xlabel('prediksi');
plt.ylabel('standardized residual');
plt.xlim([3.5,5]);
```

Setelah me-running kode pemrograman diatas, didapatkan hasil output sebagai berikut:



Gambar 27. Plot Nilai Prediksi Lama Lulusan dengan Standardized Residual

Gambar 27 terlihat bahwa data varians residual yang diwakilkan oleh standardized residualnya tidak membesar atau mengecil, sehingga dapat diartikan bahwa residual memenuhi asumsi tidak terjadi heteroskedastisitas.

Setelah melakukan pemeriksaan asumsi residual secara visual, langkah selanjutnya adalah melakukan pengujian secara statistika. Berikut adalah kode pemrograman Python pada layanan Google Colab untuk memeriksa asumsi residual dengan pengujian secara statistika.

a. Memeriksa asumsi residual berdistribusi normal

Berikut adalah kode pemrograman Python menggunakan layanan Google Colab untuk memeriksa asumsi residual berdistribusi normal dengan pengujian Jarque Bera. Kode dibawah merupakan kelanjutan dari kode pemrograman Python layanan Google Colab pada analisis regresi sederhana

```
#Pengujian Jarque
Bera untuk memeriksa asumsi residual
from statsmodels.compat import lzip
import statsmodels.formula.api as smf
import statsmodels.stats.api as sms
```

```

name = ["Jarque-Bera", "Chi^2 two-
tail prob.", "Skew", "Kurtosis"]
test = sms.jarque_bera(model.resid)
lzip(name, test)

```

Setelah me-running kode pemrograman diatas, didapatkan hasil output sebagai berikut.

```

[('Jarque-Bera', 0.6137718639092296),
 ('Chi^2 two-tail prob.', 0.7357345199195332),
 ('Skew', -0.253314975373431),
 ('Kurtosis', 2.5159095289787086)]

```

Berdasarkan Hasil output diatas maka akan dilakukan langkah-langkah pengujian Jarque-Bera

- 1) Menentukan hipotesis null dengan hipotesis alternatifnya

H_0 : Populasi berdistribusi normal

H_1 : Populasi tidak berdistribusi normal

- 2) Menentukan level of significance atau α sebesar 5% atau 0.05
- 3) Menentukan Statistik Uji

Berdasarkan output pemrograman python pengujian Jarque Bera didapatkan nilai statistik uji JB sebesar 0.6138 dan p-value sebesar 0.7357

- 4) Menentukan titik kritis pengujian

Dengan α sebesar 0.05, H_0 ditolak jika $JB > \chi^2_{(\alpha,p)}$ dimana $\chi^2_{(\alpha,p)} = \chi^2_{(0.05,3)} = 7.815$ atau p-value < 0.05

- 5) Menentukan kesimpulan hasil pengujian

Berdasarkan uji statistik didapatkan nilai JB sebesar 0.6138, dan p-value sebesar 0.7357. Dengan α sebesar 0.05 dan Tabel chi-square yaitu $\chi^2_{(\alpha,2)} = \chi^2_{(0.05,2)} = 7.815$, maka $JB < \chi^2_{(\alpha,2)}$ dan p-value > 0.05 yang dapat disimpulkan bahwa H_0 gagal ditolak artinya residual memenuhi asumsi berdistribusi normal.

b. Memeriksa asumsi residual tidak terjadi autokorelasi

Berikut adalah kode pemrograman Python menggunakan layanan Google Colab untuk memeriksa asumsi residual tidak terjadi autokorelasi dengan pengujian Durbin-Watson.

```
#Pengujian Durbin-Watson
from statsmodels.stats.stattools import durbin_watson
dw = durbin_watson(model.resid)
print(f"Durbin-Watson: {dw}")
```

Setelah me-running kode pemrograman diatas, didapatkan hasil output yaitu

Durbin-Watson: 1.261698118975068

- 1) Menentukan hipotesis null dengan hipotesis alternatifnya

H_0 : Tidak terjadi Autokorelasi

H_1 : Terjadi Autokorelasi

- 2) Menentukan level of significance atau α sebesar 5% atau 0.05

- 3) Menentukan Statistik Uji

Berdasarkan output diatas didapatkan bahwa statistic uji Durbin-Watsan adalah 1.2617

- 4) Menentukan titik kritis pengujian

Dengan α sebesar 0.05 dan $n = 30$, maka kemungkinan kesimpulan pengujian ini adalah

H_0 ditolak jika $d_{hit} < d_L(1.2837)$,

H_0 gagal ditolak jika $d_{hit} > d_U(1.5666)$

Tidak dapat menarik kesimpulan $d_L \leq d \leq d_U$

- 5) Menentukan kesimpulan hasil pengujian

Dikarenakan d_{hit} sebesar 1.262 dimana lebih kecil dari $d_L(1.2617)$ dapat disimpulkan bahwa H_0 ditolak yang artinya bahwa residual telah terjadi autokorelasi. Sehingga residual tidak memenuhi asumsi autokorelasi.

c. Memeriksa asumsi residual tidak terjadi heteroskedastisitas

Berikut adalah kode pemrograman Python menggunakan layanan Google Colab untuk memeriksa asumsi residual tidak terjadi autokorelasi dengan pengujian Breush-Pagan.

```
#Pengujian Breush-Pagan
from statsmodels.compat import lzip
import statsmodels.stats.api as sms

names = ['Lagrange multiplier statistic', 'p-value',
         'f-value', 'f p-value']
test = sms.het_breushpagan(model.resid, model.model.
                            exog)

lzip(names, test)
```

Setelah me-running kode pemrograman diatas, didapatkan hasil output yaitu

```
[['Lagrange multiplier statistic',
  1.100357485547243),
 ('p-value', 0.5768466939870686),
 ('f-value', 0.5140141801912902),
 ('f p-value', 0.603825451199743)]
```

1) Menentukan hipotesis null dengan hipotesis alternatifnya

H_0 : Tidak terjadi Heteroskedastisitas

H_1 : Terjadi Heteroskedastisitas

2) Menentukan level of significance atau α sebesar 5% atau 0.05

3) Menentukan Statistik Uji

Berdasarkan output diatas didapatkan nilai Statistik Uji Breush-Pagan sebesar 1.100 dan p-value sebesar 0.5768

4) Menentukan titik kritis pengujian

Dengan α sebesar 0.05, H_0 ditolak jika $\chi^2_{hit} > \chi^2_{\alpha,p}$ atau p-value < 0.05. $\chi^2_{0.05,2} = 5.991$

5) Menentukan kesimpulan hasil pengujian

Dikarenakan Statistik Uji Breush-Pagan sebesar 1.100 dimana lebih kecil dari $\chi^2_{0.05,2} = 5.991$ dapat disimpulkan bahwa H_0 gagal ditolak yang artinya bahwa asumsi residual tidak terjadi heteroskedastisitas.

Langkah selanjutnya adalah memeriksa apakah variabel independen memenuhi asumsi tidak terjadi multikolinieritas. Uji ini menggunakan nilai VIF atau variance Inflation Factor. Berikut adalah kode pemrograman Python menggunakan layanan Google Colab

```
from statsmodels.stats.outliers_influence import variance_inflation_factor
# Himpunan Variabel independen
X = df[['x1', 'x2']]

# VIF dataframe
vif_data = pd.DataFrame()
vif_data["feature"] = X.columns

# Menghitung VIF setiap variabel
vif_data["VIF"] = [variance_inflation_factor(X.values, i)
) for i in range(len(X.columns))]

print(vif_data)
```

Setelah me-running kode pemrograman diatas, didapatkan hasil output yaitu

	feature	VIF
0	x1	657.479364
1	x2	657.479364

Berdasarkan hasil output diatas, nilai VIF kedua variabel independen lebih besar dari 10. Hal ini mengindikasi bahwa telah terjadi multikolinieritas. Pada e-modul ini dibatasi hanya untuk memeriksa asumsi regresi. Penyelesaian dari pelanggaran ini akan dibahas pada modul edisi selanjutnya.

J. Menginterpretasikan Model

Setelah mengetahui bahwa kedua variabel berpengaruh secara singnifikan. Selanjutnya adalah menginterpretasikan model yang didapat.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\hat{y} = 10.51 - 0.03x_1 - 0.95x_2$$

Model diatas dapat diinterpretasikan setiap variabel independennya. Untuk variabel independen IQ, interpretasinya adalah setiap pertambahan satu satuan IQ yang dimiliki oleh mahasiswa akan mengurangi lama lulusan sebesar 0.03 tahun. Sedangkan untuk variabel independen IPK, interpretasinya adalah setiap pertambahan satu satuan IPK yang didapat mahasiswa akan mengurangi lama lulusan sebesar 0.95 tahun. Estimator $\hat{\beta}_0$ tidak diinterpretasikan karena tidak ada nilai variabel x yang sama dengan 0

K. Memberikan Hasil Kesimpulan

Tujuan penelitian menggunakan metode regresi linier ini adalah ingin memprediksi berapa lama lulusan mahasiswa jika mempunyai IQ sebesar 110 dan IPK semester 5 sebesar 3.25. Berikut adalah model yang didapatkan:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\hat{y} = 10.51 - 0.03(110) - 0.95(3.25)$$

$$\hat{y} = 4.12$$

Sehingga dapat disimpulkan bahwa jika mahasiswa mempunyai IQ 110 dan IPK semester 5 sebesar 3.25, maka mahasiswa lulus pada tahun ke-4 perkuliahan.

Selain itu juga berdasarkan Uji F dan Uji T dapat disimpulkan bahwa lama lulusan secara bersama-sama dan secara parsial dipengaruhi oleh IQ dan IPK mahasiswa.

V. Instruksi Tugas

Setelah memahami langkah-langkah penyelesaian permasalahan data menggunakan metode regresi berganda dengan bahasa pemrograman Python, pengguna e-modul praktikum dapat menyelesaikan tugas dengan instruksi sebagai berikut

- 1) Melakukan analisis data dengan bahasa pemrograman Python memanfaatkan layanan Google Colab
- 2) Hasil jawaban soal diketik menggunakan File Word ukuran A4 dengan format sebagai berikut:
 - a. Cover Laporan memuat Judul, Identitas pengguna, dan, Identitas Institusi
 - b. Daftar Isi, jika ada tabel dan gambar maka harus ada Daftar Tabel dan Daftar Gambar
 - c. Bab I Pendahuluan memuat latar belakang, tujuan, dan manfaat
 - d. Bab II Tinjauan Pustaka memuat teori dari metode
 - e. Bab III Metodologi data memuat dataset yang digunakan serta Langkah-langkah penyelesaian
 - f. Bab IV Hasil analisis memuat hasil penyelesaian berdasarkan soal permasalahan
 - g. Bab V Kesimpulan memuat hasil kesimpulan berdasarkan bab IV. Kesimpulan tidak boleh hasil dari copy paste
 - h. Bab VI Daftar Pustaka
- 3) Laporan dikumpulkan berbentuk Pdf.
- 4) File Notebook Google Colab hasil pemrograman Python juga ikut dikumpulkan dengan laporan
- 5) Laporan dan File Notebook Google Colab dimasukan kedalam satu folder dimana folder tersebut diberi nama identitas pengguna.

VI. Soal Regresi Linier Berganda

Kampus A setiap tahunnya menerima mahasiswa pasca sarjana. Mereka ingin mengetahui apakah kriteria-kriteria selama ini yaitu Nilai Toefl IBT dan Nilai Ujian Masuk berpengaruh terhadap peluang mahasiswa tersebut dapat

diterima di Pasca Sarjana. Selain itu juga mereka ingin memprediksi peluang mahasiswa dapat diterima di Pasca sarjana jika diketahui Skor Toefl IBT sebesar 106 dan Nilai Ujian sebesar 320. Berikut adalah data yang akan digunakan untuk menganalisis permasalahan ini.

Tabel 9. Data Persentase Kesempatan Mahasiswa Diterima di Pasca Sarjana

Negara ke-	Toefl IBT	Nilai Ujian	Peluang di Terima
1	118	337	0.92
2	107	324	0.76
3	104	316	0.72
4	110	322	0.8
5	103	314	0.65
6	115	330	0.9
7	109	321	0.75
8	101	308	0.68
9	102	302	0.5
10	108	323	0.75
11	106	325	0.75
12	111	327	0.84
13	112	328	0.78
14	109	307	0.62
15	104	311	0.61
16	105	314	0.54
17	107	317	0.66
18	106	319	0.65
19	110	318	0.63
20	102	303	0.62
21	107	312	0.64
22	114	325	0.7
23	116	328	0.94
24	119	334	0.95
25	119	336	0.97
26	120	340	0.94
27	109	322	0.76
28	98	298	0.44
29	93	295	0.46
30	99	310	0.54
31	97	300	0.65
32	103	327	0.74
33	118	338	0.91
34	114	340	0.9
35	112	331	0.94
36	110	320	0.88
37	106	299	0.64
38	105	300	0.58
39	105	304	0.65
40	108	307	0.65

Kerjakan Data tersebut berdasarkan tujuan dari Kampus A dan dalam menganalisis, langkah-langkah harus dilakukan secara lengkap dan jelas sampai kepada analisis asumsi regresi.

BAB VI

REGRESI DUMMY

I. Tujuan Pembelajaran

Tujuan pembelajaran mahasiswa setelah selesai mempelajari bab VI mengenai metode Regresi Dummy antara lain:

- a. Dapat mengetahui teori mengenai regresi dengan variabel dummy
- b. Dapat mengetahui cara penyelesaian regresi variabel dummy dengan bahasa pemrograman Python

II. Uraian Materi

Berikut adalah uraian materi bab V sehingga dapat menjawab tujuan pembelajaran yang telah ditetapkan:

1. Teori Regresi Variabel Dummy

Regresi Linier dengan variabel dummy merupakan metode regresi linier yang menyelidiki hubungan antara variabel dependen (atau respon) dan variabel independen (atau prediktor) dimana pada variabel independennya terdapat satu atau lebih variabel dengan jenis kategorik. Variabel independen dengan jenis kategorik misalkan jenis kelamin (laki-laki dan perempuan), Status pekerjaan (Bekerja dan Tidak bekerja), Shift (Pagi, Siang, dan Malam), dan lain-lain.

Misalkan manajer pabrik ingin menyelidiki apakah waktu penyelesaian produk dipengaruhi oleh Kecepatan Mesin dan Jenis Mesin. Kecepatan mesin merupakan jenis data kuantitatif, sedangkan jenis mesin merupakan jenis data kualitatif. Jenis Mesin di pabrik itu ada 2 macam yang diberi nama Mesin Tipe A dan Mesin Tipe B. Oleh karena itu sebelum menyelidiki hal tersebut, Manajer memisalkan

$$x_2 = \begin{cases} 0, & \text{Jika merupakan Mesin Tipe A} \\ 1, & \text{Jika merupakan Mesin Tipe B} \end{cases}$$

Maka model regresi berdasarkan permasalahan diatas adalah

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \quad (27)$$

Untuk menginterpretasikan model ini apabila kita mempertimbangkan Mesin Jenis A dimana $x_2 = 0$, maka model regresi menjadi

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2(0) + \varepsilon \\ y &= \beta_0 + \beta_1 x_1 + \varepsilon \end{aligned} \quad (28)$$

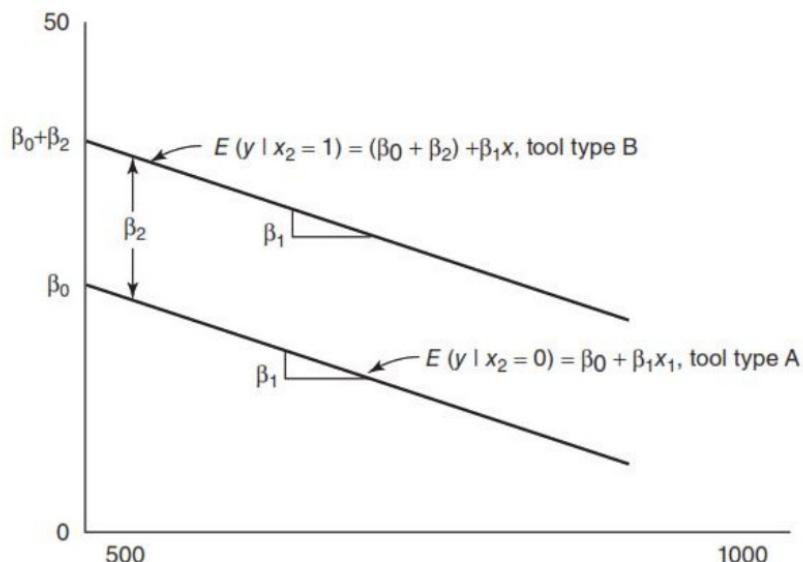
Dari model persamaan (28) didapatkan bahwa hubungan antara waktu penyelesaian produk dengan kecepatan mesin dengan tipe mesin A adalah garis lurus dengan intersept β_0 dan koefisien β_1 .

Namun jika kita mempertimbangkan Mesin Jenis B dimana $x_2 = 1$, maka model regresi menjadi

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2(1) + \varepsilon \\ y &= (\beta_0 + \beta_2) + \beta_1 x_1 + \varepsilon \end{aligned} \quad (29)$$

Dari model persamaan (29) didapatkan bahwa hubungan antara waktu penyelesaian produk dengan kecepatan mesin dengan tipe mesin B juga garis lurus dengan namun dengan intersept $\beta_0 + \beta_2$ dan koefisien β_1 .

Berdasarkan persamaan (28) dan (29), kita ketahui bahwa akan ada 2 fungsi garis regresi secara parallel yang dapat dilihat pada Gambar dibawah ini.



Gambar 28. Fungsi Garis Regresi Variabel Dummy

Berikut adalah langkah-langkah dalam menganalisis data menggunakan metode regresi dengan variabel dummy

- a) Mengetahui seberapa kuat hubungan korelasi antara variabel dependen dengan variabel dependennya menggunakan metode pearson untuk variabel independen berjenis kuantitatif dan menggunakan metode Biserial Point untuk variabel independen berjenis kategorik.
- b) Melakukan estimasi untuk mendapatkan nilai estimatornya
- c) Melakukan pengujian hipotesis Uji F untuk melihat apakah secara bersama-sama variabel independen yang lebih dari satu berpengaruh terhadap variabel dependennya,
- d) Melakukan perhitungan R-squared untuk melihat seberapa besar kesesuaian model yang didapatkan
- e) Melakukan pengujian hipotesis Uji t untuk melihat apakah secara parsial variabel independen berpengaruh terhadap variabel dependennya. Sehingga uji t ada sebanyak variabel independennya.
- f) Membuat kesimpulan dan interpretasi model
- g) Melakukan analisis asumsi residual dan asumsi multikolinieritas.
- h) Menghitung nilai prediksi dari model yang didapatkan

Dari Langkah-langkah analisis regresi dengan variabel dummy. Pada Langkah yang kedua dalam mencari korelasi hubungan antara variabel independen dan variabel dependen yang berjenis data kategorik biner yaitu menggunakan metode korelasi biserial Poin. Sama seperti metode korelasi pearson, metode ini memiliki koefisien korelasi antara -1 hingga 1. Interpretasinya pun sama dengan metode korelasi Pearson. Yang membedakannya adalah formulanya. Berikut adalah formula dari metode Biserial Poin.

$$\rho_{p,q} = \frac{M_1 - M_0}{s_n} \sqrt{pq} \quad (30)$$

Dimana

M_1 adalah rata-rata dari kelompok berkategori 1

M_0 adalah rata-rata dari kelompok berkategori 0

s_n adalah standar deviasi keseluruhan variabel

p adalah proporsi dari kelompok berkategori 0

q adalah proporsi dari kelompok berkategori 1

Untuk korelasi lebih dari 2 kategori bisa menggunakan metode Spearman atau Tau Kendall yang caranya dapat dilihat pada Bab II.

III. Rangkuman

Berikut ini adalah rangkuman materi pada Bab VI:

- a) Regresi Linier dengan variabel dummy merupakan metode regresi linier yang menyelidiki hubungan antara variabel dependen (atau respon) dan variabel independen (atau prediktor) dimana pada variabel independennya terdapat satu atau lebih variabel dengan jenis kategorik
- b) Berikut adalah langkah-langkah analisis regresi variabel dummy antara lain: membuat scatter plot variabel dependen dengan independen yang berjenis data kuantitatif, menghitung nilai korelasi variabel dependen dengan independen, mendapatkan estimator $\hat{\beta}$, melakukan Uji F, mendapatkan R-Squared, dan melakukan Uji T, memeriksa asumsi regresi, membuat kesimpulan dan interpretasi model, dan menghitung nilai prediksi dari model yang didapatkan.
- c) Metode korelasi yang digunakan untuk mengetahui seberapa kuat hubungan antara variabel dependen dan variabel independen yang berjenis data kategorik biner adalah metode korelasi Poin Biserial.

IV. Tutorial Metode

Berikut adalah tutorial metode materi bab VI yakni langkah-langkah analisis regresi variabel dummy dapat menjawab tujuan pembelajaran yang ke-2 yaitu dapat mengetahui penyelesaian regresi variabel dummy dengan bahasa pemrograman Python

1. Studi Kasus Regresi Variabel Dummy

Misalkan manajer pabrik ingin menyelidiki apakah waktu penyelesaian produk dipengaruhi oleh Kecepatan Mesin dan Jenis Mesin. Kecepatan

mesin merupakan jenis data kuantitatif, sedangkan jenis mesin merupakan jenis data kualitatif. Jenis Mesin di pabrik itu ada 2 macam yang diberi nama Mesin Tipe A dan Mesin Tipe B. Oleh karena itu sebelum menyelidiki hal tersebut, Manajer memisalkan

$$x_2 = \begin{cases} 0, & \text{Jika merupakan Mesin Tipe A} \\ 1, & \text{Jika merupakan Mesin Tipe B} \end{cases}$$

Selain itu, manajer pabrik ingin memprediksi waktu penyelesaian produk jika kecepatan mesin sebesar 990 dengan tipe mesin B. Berikut adalah data yang akan digunakan.

Tabel 10. Data Waktu Penyelesaian Produk

Data ke-1	Waktu Penyelesaian Produk (Jam)	Kecepatan Mesin (rpm)	Tipe Mesin
1	18.73	610	0
2	14.52	950	0
3	17.43	720	0
4	14.54	840	0
5	13.44	980	0
6	24.39	530	0
7	13.34	680	0
8	22.71	540	0
9	12.68	890	0
10	19.32	730	0
11	30.16	670	1
12	27.09	770	1
13	25.4	880	1
14	26.05	1000	1
15	33.49	760	1
16	35.62	590	1
17	26.07	910	1
18	36.78	650	1
19	34.95	810	1
20	43.67	500	1
27	18.73	610	1
28	14.52	950	1
29	17.43	720	1
30	14.54	840	1

2. Langkah-langkah Penyelesaian

Berikut adalah-langkah-langkah analisis regresi variabel dummy untuk menyelesaikan permasalahan pada studi kasus data Tabel 10 menggunakan bahasa pemrograman Python.

A. Menentukan Variabel Independen dan Variabel Dependen

Berdasarkan sub bab studi kasus diatas, diketahui bahwa manajer pabrik ingin menyelidiki apakah waktu penyelesaian produk dipengaruhi oleh Kecepatan Mesin dan Jenis Mesin. Dari permasalahan ini, maka dapat diketahui bahwa terdapat dua variabel independen yaitu Kecepatan Mesin (x_1) dan Jenis Mesin (x_2) dan satu variabel dependen yaitu Waktu penyelesaian Produk (y)

B. Memasukkan Data ke Notebook Google Colab

Dikarenakan jumlah data yang banyak, tidak disarankan untuk mengetik secara langsung pada Notebook Google Colab. Hal ini karena akan memakan waktu yang lama dan tidak efisien. Oleh karena itu data Tabel 10 disimpan terlebih dahulu ke bentuk File csv yang Bernama : **Waktu Produk**

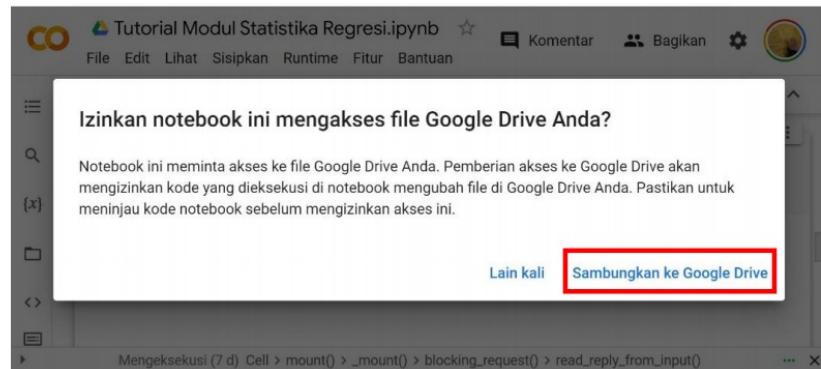
Pada e-modul Praktikum Statistika Regresi ini, dalam memanggil data bentuk File csv terlebih dahulu disimpan di Folder Google Drive masing-masing pengguna e-modul. Penulis menyimpan File csv pada Folder yang bernama : **Dataset eModul**.

Berikut adalah kode pemrograman python pada notebook Google Colab untuk memasukan data ke Notebook Google Colab

- 1) Menghubungkan Google Drive dengan Notebook Google Colab

```
#CONNECT GOOGLE DRIVE
from google.colab import drive
drive.mount('/content/drive')
```

Setelah me-running kode diatas akan muncul Gambar yang dapat dilihat pada Gambar 29 dan klik **Sambungkan ke Google Drive**. Kemudian akan muncul Gambar 30 untuk memilih akun Google Drive tempat meyimpan dataset kita: **Klik Akun → Izinkan**. Setelah itu dataset di Google Drive sudah dapat terbaca oleh Notebook Google Colab yang ditandai dengan kode : Mounted at /content/drive



Gambar 29. Izin Mengakses Google Drive Data Regresi Berganda



Gambar 30. Memilih Akun Google Drive untuk Data Regresi Berganda

2) Membuat dataframe data menggunakan library Pandas

Setelah data terhubung dari Google Drive ke Notebook Google Colab, Selanjutnya adalah menyusun dataset kita menjadi data frame dengan library Pandas. Kode pemrograman python-nya adalah

```
# Memanggil dataset
import pandas as pd
df = pd.read_csv("drive/MyDrive/Dataset eModul/Waktu
Produk.csv")
df.head()
```

Kode diatas akan menghasilkan Output yaitu



	Waktu	Kecepatan	Tipe Mesin
0	18.73	610	0
1	14.52	950	0
2	17.43	720	0
3	14.54	840	0
4	13.44	980	0

Gambar 31. Output Data Waktu Penyelesaian Mesin

Diketahui bahwa untuk mengefisienkan penamaan kode pemrograman python, maka nama variabel Gambar 31 diganti dari **Kecepatan** menjadi x_1 dan **Tipe Mesin** menjadi x_2 , serta **Waktu** diganti menjadi y . Berikut adalah kode pemrogramannya

```
#Mengganti nama variabel
df.rename(columns={'Kecepatan':'x1','Tipe Mesin':'x2',
,'Waktu ':'y'}, inplace=True)
df.head()
```

Kode diatas akan menghasilkan output yaitu



	y	x1	x2
0	18.73	610	0
1	14.52	950	0
2	17.43	720	0
3	14.54	840	0
4	13.44	980	0

Gambar 32. Mengganti Nama Variabel di Data Waktu Penyelesaian Produk

C. Mengetahui Seberapa Kuat Hubungan Variabel Dependensi dan Independen Menggunakan Metode Korelasi

Langkah selanjutnya adalah mengetahui seberapa kuat variabel x_1 dan x_2 dapat mempengaruhi variabel dependennya. Metode yang digunakan adalah metode korelasi Pearson untuk berjenis data kuantitatif dan metode Biserial Poin untuk jenis data kualitatif biner. Berikut adalah kode pemrograman python yang digunakan untuk mengetahui seberapa kuat hubungan variabel Kecepatan Mesin (x_1) terhadap variabel Waktu Penyelesaian Produk (y).

```
#Mengitung nilai korelasi pearson
from scipy.stats import pearsonr
# Convert dataframe into series
list1 = df['x1']
list2 = df['y']
corr, _ = pearsonr(list1, list2)
print('Koefisien Pearson: %.5f' % corr)
```

Hasil output dari kode diatas adalah

```
Koefisien Pearson: -0.43131
```

Berdasarkan nilai korelasi Pearson sebesar -0.43131 maka dapat disimpulkan bahwa variable kecepatan mesin dengan variabel Waktu penyelesaian produk memiliki hubungan korelasi negatif yang lemah karena memiliki nilai koefisien korelasi kurang dari -0.5

Berikut adalah kode pemrograman python yang digunakan untuk mengetahui seberapa kuat hubungan variabel Tipe Mesin (x_2) terhadap variabel lama lulus (y) menggunakan metode Poin Biserial

```
import scipy.stats as stats
#menghitung Korelasi Biserial Poin Tipe Mesin VS Waktu Penyelesaian Produk
stats.pointbiserialr(df['x2'], df['y'])
```

Hasil output dari kode diatas adalah

```
PointbiserialrResult (correlation=0.8348763455280519,
pvalue=4.681443394353966e-06)
```

Berdasarkan nilai korelasi Biserial Poin sebesar 0.83488 maka dapat disimpulkan bahwa variabel Tipe Mesin dengan variabel Waktu

Penyelesaian Produk memiliki hubungan korelasi positif yang kuat karena memiliki nilai koefisien korelasi lebih dari 0.8.

D. Membentuk Model Regresi Dummy

Langkah selanjutnya memodelkan dataset dengan menggunakan analisis regresi dummy. Kode pemrograman python adalah sebagai berikut.

```
#Memodelkan dengan Regresi Linier Berganda
import numpy as np
import statsmodels.formula.api as smf
df_with_dummies = pd.get_dummies(data=df, columns=['x2'])
x = df_with_dummies[['x1','x2_0','x2_1']]
y = df_with_dummies['y']
reg_mod = 'y ~ x1+x2_1'
model = smf.ols(formula=reg_mod, data=df_with_dummies).fit()
print_model = model.summary()
print(print_model)
```

Hasil output dari Kode pemrograman Python diatas adalah

```
OLS Regression Results
=====
Dep. Variable: y R-squared: 0.900
Model: OLS Adj. R-squared: 0.889
Method: Least Squares F-statistic: 76.75
Date: Thu, 17 Nov 2022 Prob (F-statistic): 3.09e-09
Time: 15:13:20 Log-Likelihood: -48.987
No. Observations: 20 AIC: 104.0
Df Residuals: 17 BIC: 107.0
Df Model: 2
Covariance Type: nonrobust
=====
            coef    std err        t      P>|t|      [0.025      0.975]
-----
Intercept  36.9856   3.510     10.536      0.000     29.579     44.392
x1         -0.0266   0.005     -5.887      0.000     -0.036     -0.017
x2_1       15.0043   1.360     11.035      0.000     12.136     17.873
=====
Omnibus: 0.346 Durbin-Watson: 1.322
Prob(Omnibus): 0.841 Jarque-Bera (JB): 0.497
Skew: -0.204 Prob(JB): 0.780
Kurtosis: 2.344 Cond. No. 3.96e+03
=====
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.96e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Gambar 33. Output Model Regresi Dummy

Berdasarkan Gambar 35 Jika Tipe Mesin adalah A atau 0, maka model regresi yang didapat adalah

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \\ \hat{y} &= 36.986 - 0.03x_1 + 15.00(0) \\ \hat{y} &= 36.986 - 0.03x_1\end{aligned}\tag{31}$$

Jika Tipe Mesin adalah B atau 1, maka model regresi yang didapat adalah

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \\ \hat{y} &= 36.986 - 0.03x_1 + 15.00(1) \\ \hat{y} &= (36.986 + 15.00) - 0.03x_1 \\ \hat{y} &= 51.986 - 0.03x_1\end{aligned}\tag{32}$$

Dari Persamaan (31) dan (32) kita akan dapat menghasilkan nilai prediksi dan residualnya. Kode pemrograman python untuk menghasilkan nilai prediksi \hat{y} adalah

```
prediksi = model.predict(x)
print(prediksi.head())
```

Output hasil kode pemrograman diatas antara lain

```
0    20.755191
1    11.708732
2    17.828395
3    14.635528
4    10.910515
dtype: float64
```

Adapun kode pemrograman python untuk menghasilkan nilai residual e_i adalah

```
residual=model.resid
print(residual.head())
```

Output hasil kode pemrograman diatas antara lain

```
0    -2.025191
1     2.811268
2    -0.398395
3    -0.095528
```

```
4      2.529485
dtype: float64
```

E. Melakukan Uji F

Setelah mendapatkan model regresi dummy, langkah selanjutnya adalah mengetahui apakah secara bersama-sama ada minimal satu variabel independen yang berpengaruh terhadap variabel dependennya. Untuk itu diperlukan suatu Pengujian Hipotesis Uji F dengan langkah-langkah sebagai berikut.

6. Menentukan hipotesis null dengan hipotesis alternatifnya

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

$$H_1 : \text{minimal ada satu } \beta_j \neq 0, \text{ dimana } j = 1, 2$$

7. Menentukan level of significance atau α sebesar 5% atau 0.05

8. Menentukan Uji Statistik F

Pada Gambar 35, didapatkan bahwa Nilai F -statistic yang diperoleh adalah 76.75 atau p-value yang didapat sebesar 3.09e-09

9. Menentukan titik kritis pengujian

Dengan α sebesar 0,05, H_0 ditolak jika $F_{hit} > F_{tabel}$ dimana $F_{tabel} = F_{0.05,2,17} = 3.5915$ atau p-value < 0,05.

10. Menentukan kesimpulan hasil pengujian

Dengan α sebesar 0,05, nilai $F_{hit} = 76.75$ lebih besar dibandingkan $F_{tabel} = F_{0.05,2,17} = 3.5915$ dan juga nilai p-value yang didapatkan sebesar 0.00 yang kurang dari 0,05. Berdasarkan hal itu maka dapat disimpulkan bahwa H_0 ditolak yang artinya minimal ada satu variabel independen yang berpengaruh terhadap variabel dependennya.

F. Mendapatkan Nilai R-Squared

Langkah selanjutnya adalah mengetahui nilai proporsi dari total varians yang dijelaskan oleh model regresi atau dikenal sebagai nilai R-Squared. Pada analisis regresi dummy, karena memiliki dua

variabel independen maka lebih baik menggunakan R-Squared Adjusted untuk melihat seberapa besar variabel dependen dapat dijelasakan oleh variabel independennya. Berdasarkan Gambar 35, didapatkan nilai R-Squared Adjusted sebesar 0.889 yang artinya bahwa 88,9 % waktu penyelesaian produk oleh kecepatan mesin dan tipe mesin dan sisanya 11,1% dipengaruhi oleh variabel lainnya yang tidak diketahui atau tidak diteliti.

G. Melakukan Uji T

Langkah selanjutnya adalah melihat apakah variabel independen berpengaruh secara signifikan atau tidak terhadap variabel dependennya. Dikarenakan variabel independen berjumlah dua, maka Uji T akan dilakukan sebanyak dua kali. Langkah-langkah Uji T untuk variabel IQ adalah sebagai berikut.

1. Menentukan hipotesis null dengan hipotesis alternatifnya

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

2. Menentukan level of significance atau α sebesar 5% atau 0.05

3. Menentukan Uji Statistik t-hitung

Pada Gambar 24, didapatkan bahwa Nilai t -statistic untuk variabel Kecepatan Mesin adalah -5.887 atau p-value yang didapat sebesar 0.000

4. Menentukan titik kritis pengujian

Dengan α sebesar 0.05, H_0 ditolak jika $|t_{hit}| > t_{tabel}$ dimana $t_{tabel} = t_{\frac{\alpha}{2}, n-3} = t_{0.025; 17} = 2.1098$

5. Menentukan kesimpulan hasil pengujian

Dengan α sebesar 0,05, nilai $|t_{hit}| = 5.887$ lebih besar dibandingkan $t_{tabel} = 2.1098$ dan juga nilai p-value yang didapatkan sebesar 0.000 yang kurang dari 0,05. Berdasarkan hal itu maka dapat disimpulkan bahwa H_0 ditolak yang artinya bahwa

variabel kecepatan mesin berpengaruh signifikan terhadap waktu penyelesaian produk

Berikut adalah langkah-langkah Uji T untuk variabel IPK apakah variabel tersebut berpengaruh signifikan terhadap variabel lama lulus.

1. Menentukan hipotesis null dengan hipotesis alternatifnya

$$H_0 : \beta_2 = 0$$

$$H_1 : \beta_2 \neq 0$$

2. Menentukan level of significance atau α sebesar 5% atau 0.05

3. Menentukan Uji Statistik t-hitung

Pada Gambar 24, didapatkan bahwa Nilai $t\text{-statistic}$ untuk variabel Tipe Mesin adalah 11.035 atau p-value yang didapat sebesar 0.000

4. Menentukan titik kritis pengujian

Dengan α sebesar 0.05, H_0 ditolak jika $|t_{hit}| > t_{tabel}$ dimana

$$t_{tabel} = t_{\frac{\alpha}{2}, n-3} = t_{0.025; 27} = 2.1098$$

5. Menentukan kesimpulan hasil pengujian

Dengan α sebesar 0,05, nilai $|t_{hit}| = 11.035$ lebih besar dibandingkan $t_{tabel} = 2.1098$ dan juga nilai p-value yang didapatkan sebesar 0.000 yang kurang dari 0,05. Berdasarkan hal itu maka dapat disimpulkan bahwa H_0 ditolak yang artinya bahwa variabel Tipe Mesin berpengaruh signifikan terhadap waktu penyelesaian produk

H. Uji Asumsi Regresi

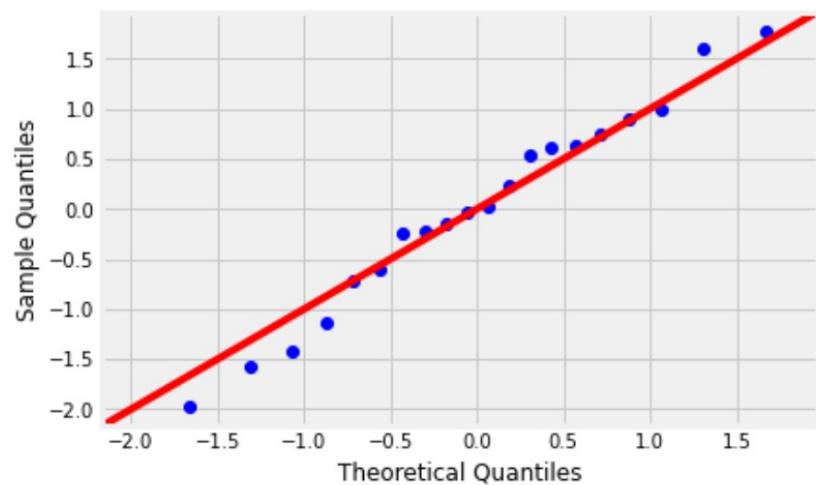
Setelah mendapatkan residual pada model regresi linier dummy, langkah selanjutnya adalah memeriksa asumsi residual dan asumsi tidak terjadinya multikolinieritas. Berikut adalah langkah-langkah menganalisis asumsi residual menggunakan bahasa pemrograman Python pada layanan Google Colab menggunakan suatu plot

- a. Memeriksa asumsi residual berdistribusi normal

Berikut adalah kode pemrograman Python menggunakan layanan Google Colab untuk memeriksa asumsi residual berdistribusi normal dengan Q-Q Plot.

```
#Menggambar Plot QQ menggunakan package statmodels
import scipy.stats as stats
fig = sm.qqplot(residual, stats.t, fit=True, line="45")
plt.show()
```

Setelah me-running kode pemrograman diatas, didapatkan hasil output sebagai berikut:



Gambar 34. Plot Q-Q data Residual Waktu Penyelesaian Produk

Gambar 36 terlihat bahwa distribusi data residual mendekati garis linier. Sehingga dapat disimpulkan bahwa data residual tersebut memenuhi asumsi berdistribusi normal.

- c. Memeriksa asumsi residual tidak terjadi autokorelasi

Berikut adalah kode pemrograman Python menggunakan layanan Google Colab untuk memeriksa asumsi residual tidak terjadi autokorelasi.

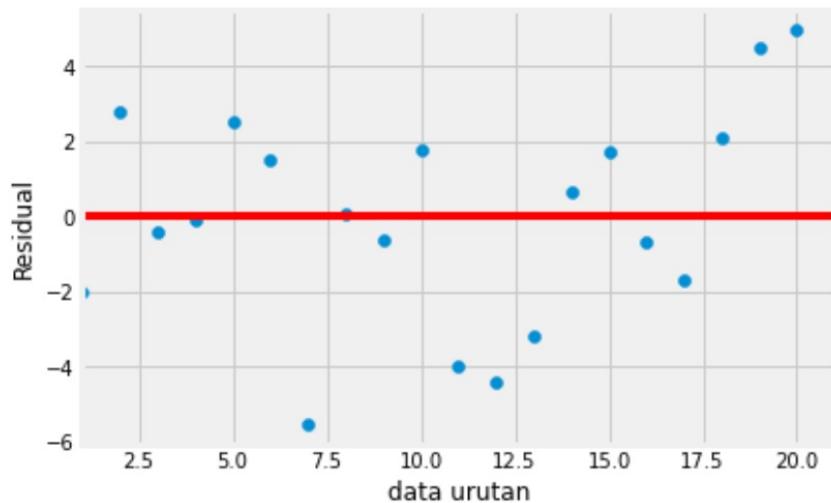
```
#Plot memeriks asumsi residual tidak terjadi autokore
lasi
urutan_pengamatan=pd.Series(range(1,31))
```

```

plt.scatter(urutan_pengamatan, residual);
plt.axhline(0, color='red')
plt.xlabel('data urutan');
plt.ylabel('Residual');
plt.xlim([1,21]);

```

Setelah me-running kode pemrograman diatas, didapatkan hasil output sebagai berikut:



Gambar 35. Plot Residual Waktu Penyelesaian Produk

Gambar 37 terlihat bahwa distribusi data residual terhadap plot tidak membentuk pola apapun. Sehingga dapat disimpulkan bahwa data residual memenuhi asumsi tidak terjadinya autokorelasi.

d. Memeriksa asumsi residual tidak terjadi heteroskedastisitas

Berikut adalah kode pemrograman Python menggunakan layanan Google Colab untuk memeriksa asumsi residual tidak terjadi heteroskedastisitas.

```

#Plot memeriksa asumsi residual tidak terjadi heteros
kedastisitas
influence = model.get_influence()
#menentukan standardized residualnya
std_residual = influence.resid_studentized_internal
plt.scatter(prediksi, std_residual);

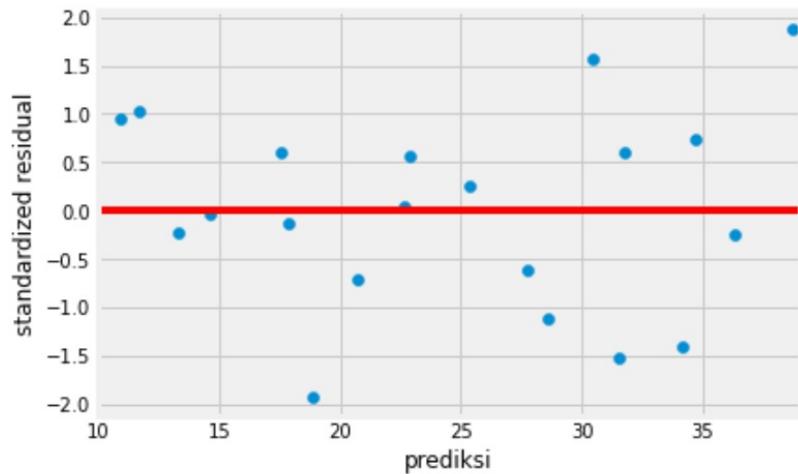
```

```

plt.axhline(0, color='red')
plt.xlabel('prediksi');
plt.ylabel('standardized residual');
plt.xlim([10,39]);

```

Setelah me-running kode pemrograman diatas, didapatkan hasil output sebagai berikut:



Gambar 36. Plot Nilai Waktu Penyelesaian Produk dengan Standardized Residual

Gambar 38 terlihat bahwa data varians residual yang diwakilkan oleh standardized residualnya tidak membentuk pol apapun, sehingga dapat diartikan bahwa residual memenuhi asumsi tidak terjadi heteroskedastisitas.

Setelah melakukan pemeriksaan asumsi residual secara visual, langkah selanjutnya adalah melakukan pengujian secara statistika. Berikut adalah kode pemrograman Python pada layanan Google Colab untuk memeriksa asumsi residual dengan pengujian secara statistika.

d. Memeriksa asumsi residual berdistribusi normal

Berikut adalah kode pemrograman Python menggunakan layanan Google Colab untuk memeriksa asumsi residual berdistribusi normal dengan pengujian Jarque Bera. Kode dibawah merupakan

kelanjutan dari kode pemrograman Python layanan Google Colab pada analisis regresi sederhana

```
#Pengujian Jarque
Bera untuk memeriksa asumsi residual
from statsmodels.compat import lzip
import statsmodels.formula.api as smf
import statsmodels.stats.api as sms
name = ["Jarque-Bera", "Chi^2 two-
tail prob.", "Skew", "Kurtosis"]
test = sms.jarque_bera(model.resid)
lzip(name, test)
```

Setelah me-running kode pemrograman diatas, didapatkan hasil output sebagai berikut.

```
[('Jarque-Bera', 0.4971653524602619),
 ('Chi^2 two-tail prob.', 0.7799053785330367),
 ('Skew', -0.20405165821810348),
 ('Kurtosis', 2.344217943165503)]
```

Berdasarkan Hasil output diatas maka akan dilakukan langkah-langkah pengujian Jarque-Bera

- 1) Menentukan hipotesis null dengan hipotesis alternatifnya
 H_0 : Populasi berdistribusi normal
 H_1 : Populasi tidak berdistribusi normal
- 2) Menentukan level of significance atau α sebesar 5% atau 0.05
- 3) Menentukan Statistik Uji

Berdasarkan output pemrograman python pengujian Jarque Bera didapatkan nilai statistik uji JB sebesar 0.4972 dan p-value sebesar 0.7799

- 4) Menentukan titik kritis pengujian
 Dengan α sebesar 0.05, H_0 ditolak jika $JB > \chi^2_{(\alpha,p)}$ dimana $\chi^2_{(0.05,3)} = 7.815$ atau p-value < 0.05
- 5) Menentukan kesimpulan hasil pengujian

Berdasarkan uji statistik didapatkan nilai JB sebesar 0.4972, dan p-value sebesar 0.7357. Dengan α sebesar 0.05 dan Tabel chi-square yaitu $\chi^2_{(\alpha,2)} = \chi^2_{(0.05,2)} = 7.815$, maka $JB < \chi^2_{(\alpha,2)}$ dan p-value > 0.05 yang dapat disimpulkan bahwa H_0 gagal ditolak artinya residual memenuhi asumsi berdistribusi normal.

e. Memeriksa asumsi residual tidak terjadi autokorelasi

Berikut adalah kode pemrograman Python menggunakan layanan Google Colab untuk memeriksa asumsi residual tidak terjadi autokorelasi dengan pengujian Durbin-Watson.

```
#Pengujian Durbin-Watson
from statsmodels.stats.stattools import durbin_watson
dw = durbin_watson(model.resid)
print(f"Durbin-Watson: {dw}")
```

Setelah me-running kode pemrograman diatas, didapatkan hasil output yaitu

Durbin-Watson: 1.3222499319171725

- 1) Menentukan hipotesis null dengan hipotesis alternatifnya
 H_0 : Tidak terjadi Autokorelasi
 H_1 : Terjadi Autokorelasi
- 2) Menentukan level of significance atau α sebesar 5% atau 0.05
- 3) Menentukan Statistik Uji

Berdasarkan output diatas didapatkan bahwa statistic uji Durbin-Watsan adalah 1.3222

- 4) Menentukan titik kritis pengujian

Dengan α sebesar 0.05 dan $n = 20$, maka kemungkinan kesimpulan pengujian ini adalah

H_0 ditolak jika $d_{hit} < d_L(1.1004)$,
 H_0 gagal ditolak jika $d_{hit} > d_U(1.5367)$
Tidak dapat menarik kesimpulan $d_L \leq d \leq d_U$

- 5) Menentukan kesimpulan hasil pengujian

Dikarenakan d_{hit} sebesar 1.3222 dimana diantara $d_L \leq d \leq d_U$ maka tidak dapat disimpulkan dengan metode durbin Watson bahwa residual terjadi autokorelasi atau tidak.

f. Memeriksa asumsi residual tidak terjadi heteroskedastisitas

Berikut adalah kode pemrograman Python menggunakan layanan Google Colab untuk memeriksa asumsi residual tidak terjadi autokorelasi dengan pengujian Breush-Pagan.

```
#Pengujian Breush-Pagan
from statsmodels.compat import lzip
import statsmodels.stats.api as sms

names = ['Lagrange multiplier statistic', 'p-value',
         'f-value', 'f p-value']
test = sms.het_breushpagan(model.resid, model.model.
                            exog)

lzip(names, test)
```

Setelah me-running kode pemrograman diatas, didapatkan hasil output yaitu

```
[('Lagrange multiplier statistic',
  1.9587408768777936),
 ('p-value', 0.3755474546851591),
 ('f-value', 0.9228456472931548),
 ('f p-value', 0.4164023894220722)]
```

1) Menentukan hipotesis null dengan hipotesis alternatifnya

H_0 : Tidak terjadi Heteroskedastisitas

H_1 : Terjadi Heteroskedastisitas

2) Menentukan level of significance atau α sebesar 5% atau 0.05

3) Menentukan Statistik Uji

Berdasarkan output diatas didapatkan nilai Statistik Uji Breush-Pagan sebesar 1.9587 dan p-value sebesar 0.3755

4) Menentukan titik kritis pengujian

Dengan α sebesar 0.05, H_0 ditolak jika $\chi^2_{hit} > \chi^2_{\alpha,p}$ atau p-value < 0.05 . $\chi^2_{0.05,2} = 5.991$

5) Menentukan kesimpulan hasil pengujian

Dikarenakan Statistik Uji Breush-Pagan sebesar 1.9587 dimana lebih kecil dari $\chi^2_{0.05,2} = 5.991$ dapat disimpulkan bahwa H_0 gagal ditolak yang artinya bahwa asumsi residual tidak terjadi heteroskedastisitas.

Langkah selanjutnya adalah memeriksa apakah variabel independen memenuhi asumsi tidak terjadi multikolinieritas. Uji ini menggunakan nilai VIF atau Variance Inflation Factor. Berikut adalah kode pemrograman Python menggunakan layanan Google Colab

```
from statsmodels.stats.outliers_influence import variance_inflation_factor
# Himpunan Variabel Independen
X = df[['x1', 'x2']]

# VIF dataframe
vif_data = pd.DataFrame()
vif_data["feature"] = X.columns

# Menghitung VIF setiap variabel
vif_data["VIF"] = [variance_inflation_factor(X.values, i)
) for i in range(len(X.columns))]

print(vif_data)
```

Setelah me-running kode pemrograman diatas, didapatkan hasil output yaitu

feature	VIF
0 x1	1.942447
1 x2_1	1.942447

Berdasarkan hasil output diatas, nilai VIF kedua variabel independen lebih kecil dari 10. Hal ini mengindikasi bahwa variabel independen tidak terjadi multikolinieritas.

I. Menginterpretasikan Model

Setelah mengetahui bahwa kedua variabel berpengaruh secara signifikan. Selanjutnya adalah menginterpretasikan model yang didapat. Interpretasi model regresi dengan tipe mesin jenis A adalah

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \\ \hat{y} &= 36.986 - 0.03x_1 + 15.00(0) \\ \hat{y} &= 36.986 - 0.03x_1\end{aligned}$$

Model diatas dapat diinterpretasikan bahwa setiap pertambahan satu satuan kecepatan mesin akan mengurangi waktu penyelesaian produk sebesar 0.03 jam. Estimator $\hat{\beta}_0$ tidak diinterpretasikan karena tidak ada nilai variabel x yang sama dengan 0

Interpretasi model regresi dengan tipe mesin jenis B adalah

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \\ \hat{y} &= 36.986 - 0.03x_1 + 15.00(1) \\ \hat{y} &= (36.986 + 15.00) - 0.03x_1 \\ \hat{y} &= 51.986 - 0.03x_1\end{aligned}$$

Model diatas dapat diinterpretasikan bahwa setiap pertambahan satu satuan kecepatan mesin akan mengurangi waktu penyelesaian produk sebesar 0.03 jam. Estimator $\hat{\beta}_0$ tidak diinterpretasikan karena tidak ada nilai variabel x yang sama dengan 0

J. Memberikan Hasil Kesimpulan

Tujuan penelitian menggunakan metode regresi variabel dummy ini adalah ingin waktu penyelesaian produk jika kecepatan mesin sebesar 990 dengan tipe mesin B. Berikut adalah model yang didapatkan:

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \\ \hat{y} &= 36.986 - 0.03x_1 + 15.00(1) \\ \hat{y} &= (36.986 + 15.00) - 0.03x_1\end{aligned}$$

$$\begin{aligned}\hat{y} &= 51.986 - 0.03x_i \\ \hat{y} &= 51.986 - 0.03(990) \\ \hat{y} &= 22.286\end{aligned}$$

Sehingga dapat disimpulkan bahwa jika Mesin Tipe B dengan kecepatan sebesar 990 maka waktu penyelesaian produk diprediksi menjadi 22.286 jam

Selain itu juga berdasarkan Uji F dan Uji T dapat disimpulkan bahwa waktu penyelesaian produk secara bersama-sama dan secara parsial dipengaruhi oleh kecepatan mesin dan Tipe mesinya.

V. Instruksi Tugas

Setelah memahami langkah-langkah penyelesaian permasalahan data menggunakan metode regresi variabel dummy dengan bahasa pemrograman Python, pengguna e-modul praktikum dapat menyelesaikan tugas dengan instruksi sebagai berikut

- 1) Melakukan analisis data dengan bahasa pemrograman Python memanfaatkan layanan Google Colab
- 2) Hasil jawaban soal diketik menggunakan File Word ukuran A4 dengan format sebagai berikut:
 - a. Cover Laporan memuat Judul, Identitas pengguna, dan, Identitas Institusi
 - b. Daftar Isi, jika ada tabel dan gambar maka harus ada Daftar Tabel dan Daftar Gambar
 - c. Bab I Pendahuluan memuat latar belakang, tujuan, dan manfaat
 - d. Bab II Tinjauan Pustaka memuat teori dari metode
 - e. Bab III Metodologi data memuat dataset yang digunakan serta Langkah-langkah penyelesaian
 - f. Bab IV Hasil analisis memuat hasil penyelesaian berdasarkan soal permasalahan
 - g. Bab V Kesimpulan memuat hasil kesimpulan berdasarkan bab IV. Kesimpulan tidak boleh hasil dari copy paste
 - h. Bab VI Daftar Pustaka
- 3) Laporan dikumpulkan berbentuk Pdf.

- 4) File Notebook Google Colab hasil pemrograman Python juga ikut dikumpulkan dengan laporan
- 5) Laporan dan File Notebook Google Colab dimasukan kedalam satu folder dimana folder tersebut diberi nama identitas pengguna.

V. Soal Regresi Variabel Dummy

Suatu Showroom mobil ingin mengetahui apakah ukuran mesin dan desain mesin berpengaruh terhadap harga jual mobil. Adapun desain mobil merupakan jenis data kualitatif yang terdiri dari Standar dan Turbo. Dimisalkan standar =0 dan turbo = 1. Selain itu Showroom tersebut ingin memprediksi berapa harga jual mobil jika desain mobilnya standar dengan ukuran mesin 94. Berikut adalah data yang akan digunakan untuk menganalisis permasalahan ini.

Tabel 11. Data Harga Jual Mobil

Negara ke-	Ukuran Mesin	Desian Mobil	Harga Jual Mobil
1	130	1	13495
2	130	1	16500
3	152	1	16500
4	109	1	13950
5	136	1	17450
6	136	1	15250
7	136	1	17710
8	136	1	18920
9	131	1	23875
10	108	1	16430
11	108	1	16925
12	164	1	20970
13	164	1	21105
14	164	1	24565
15	209	1	30760
16	209	1	41315
17	209	1	36880
18	61	0	5151
19	90	0	6295
20	90	0	6575
21	90	0	5572
22	90	0	6377
23	98	0	7957
24	90	0	6229
25	90	0	6692
26	90	0	7609
27	98	0	8558
28	122	0	8921
29	156	0	12964
30	92	0	6479

Negara ke-	Ukuran Mesin	Desian Mobil	Harga Jual Mobil
31	92	0	6855
32	79	0	5399
33	92	0	6529
34	92	0	7129
35	92	0	7295
36	92	0	7295
37	110	0	7895
38	110	0	9095
39	110	0	8845
40	110	1	10295

Kerjakan Data tersebut berdasarkan tujuan dari Showroom dan dalam menganalisis, langkah-langkah harus dilakukan secara lengkap dan jelas sampai kepada analisis asumsi regresi.

BAB VII

REGRESI POLYNOMIAL

I. Tujuan Pembelajaran

Tujuan pembelajaran mahasiswa setelah selesai mempelajari bab VII mengenai metode Regresi Polynomial antara lain:

- c. Dapat mengetahui teori mengeani regresi dengan variabel polynomial
- d. Dapat mengetahui cara penyelesaian regresi variabel polynomial dengan bahasa pemrograman Python

II. Uraian Materi

Berikut adalah uraian materi bab VII sehingga dapat menjawab tujuan pembelajaran yang telah ditetapkan:

1. Teori Regresi Variabel Polynomial

Regresi variabel polynomial merupakan metode yang menyelidiki hubungan non-linier antara variabel dependen (atau respon) dan variabel independen (atau prediktor) dengan fungsi regresi polynomial. Model regresi polynomial dengan satu variabel independen adalah:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_h x^h + \varepsilon \quad (33)$$

Dimana h adalah orde dari polynomial. Untuk orde ke-2 disebut kuadratik, orde ke-3 disebut kubik, orde ke-4 disebut kuartik, dan seterusnya. Meskipun hubungannya non-linier namun regresi ini masih dianggap regresi linier karena masih linier dalam koefisien regresinya. Untuk menemukan orde polynomial yang tepat, kita dapat menggunakan metode **Forward Selection** dan **Backward Selection**. **Forward Selection** merupakan metode yang terus meningkatkan orde sampai cukup signifikan untuk menemukan model yang terbaik. Sedangkan **Backward Selection** merupakan metode yang terus mengurangi orde sampai cukup signifikan untuk menemukan model yang terbaik. Pada e-

modul ini akan menggunakan Metode Forward Selection untuk mencari order model polynomial yang tepat.

Berikut adalah langkah-langkah dalam menganalisis data menggunakan metode regresi dengan variabel polynomial

- a) Memodelkan data dengan menggunakan regresi polynomial orde pertama atau linier
- b) Melihat apakah dengan Uji F, variabel independennya berpengaruh atau tidak dengan variabel dependennya.
- c) Jika tidak signifikan, dilanjutkan ke orde yang lebih tinggi hingga didapatkan variabel independen signifikan terhadap variabel dependen.
- d) Melakukan perhitungan R-squared untuk melihat seberapa besar kesesuaian model terbaiknya.
 - a) Melakukan analisis asumsi residual
 - b) Membuat kesimpulan dan interpretasi model

III. Rangkuman

Berikut ini adalah rangkuman materi pada Bab VII:

- a) Regresi variabel polynomial merupakan metode yang menyelidiki hubungan non-linier antara variabel dependen (atau respon) dan variabel independen (atau prediktor) dengan fungsi regresi polynomial
- b) Berikut adalah langkah-langkah analisis regresi variabel dummy antara lain: memodelkan data dengan menggunakan regresi polynomial orde pertama atau linier, Melihat apakah dengan Uji F variabel independennya berpengaruh atau tidak dengan variabel dependennya, Jika tidak signifikan, dilanjutkan ke orde yang lebih tinggi hingga didapatkan variabel independen signifikan terhadap variabel dependen, melakukan perhitungan R-squared untuk melihat seberapa besar kesesuaian model terbaiknya memeriksa asumsi regresi, membuat kesimpulan dan interpretasi model, dan menghitung nilai prediksi dari model yang didapatkan.

IV. Tutorial Metode

Berikut adalah tutorial metode materi bab VII yakni langkah-langkah analisis regresi variabel polynomial dengan bahasa pemrograman Python.

1. Studi Kasus Regresi Polynomial

Misalkan peneliti di Laboratorium ingin menyelidiki apakah hasil percobaannya dipengaruhi oleh suhu temperature. Datanya adalah sebagai berikut.

Tabel 12. Data Percobaan

Data ke-1	Suhu Temperatur	Hasil Percobaan
1	50	3.3
2	50	2.8
3	50	2.9
4	70	2.3
5	70	2.6
6	70	2.1
7	80	2.5
8	80	2.9
9	80	2.4
10	90	3
11	90	3.1
12	90	2.8
13	100	3.3
14	100	3.5
15	100	3

2. Langkah-langkah Penyelesaian

Berikut adalah langkah-langkah analisis regresi polynomial untuk menyelesaikan permasalahan pada studi kasus data Tabel 12 menggunakan bahasa pemrograman Python.

A. Menentukan Variabel Independen dan Variabel Dependental

Sesuai studi kasus diatas, diketahui bahwa variabel dependen adalah hasil percobaan (y) dan variabel indepennya adalah temperature (x).

B. Memasukkan Data ke Notebook Google Colab

Tabel 12 disimpan terlebih dahulu ke bentuk File csv yang Bernama : **Data Percobaan**

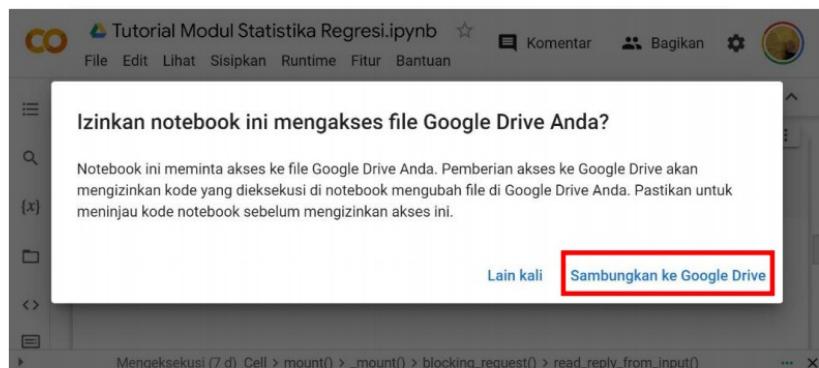
Pada e-modul Praktikum Statistika Regresi ini, dalam memanggil data bentuk File csv terlebih dahulu disimpan di Folder Google Drive masing-masing pengguna e-modul. Penulis menyimpan File csv pada Folder yang bernama : **Dataset eModul**.

Berikut adalah kode pemrograman python pada notebook Google Colab untuk memasukan data ke Notebook Google Colab

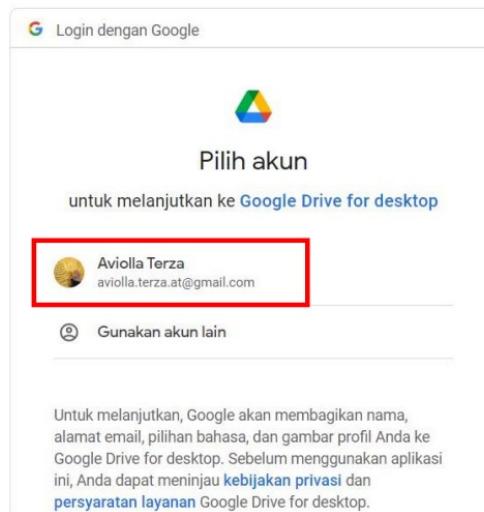
3) Menghubungkan Google Drive dengan Notebook Google Colab

```
#CONNECT GOOGLE DRIVE
from google.colab import drive
drive.mount('/content/drive')
```

Setelah me-running kode diatas akan muncul Gambar yang dapat dilihat pada Gambar 29 dan klik **Sambungkan ke Google Drive**. Kemudian akan muncul Gambar 30 untuk memilih akun Google Drive tempat meyimpan dataset kita: **Klik Akun → Izinkan**. Setelah itu dataset di Google Drive sudah dapat terbaca oleh Notebook Google Colab yang ditandai dengan kode : Mounted at /content/drive



Gambar 37. Izin Mengakses Google Drive Data Regresi Berganda



Gambar 38. Memilih Akun Google Drive untuk Data Regresi Berganda

4) Membuat dataframe data menggunakan library Pandas

Setelah data terhubung dari Google Drive ke Notebook Google Colab, Selanjutnya adalah menyusun dataset kita menjadi data frame dengan library Pandas. Kode pemrograman python-nya adalah

```
# Memanggil dataset
import pandas as pd
df = pd.read_csv("drive/MyDrive/Dataset eModul/Data percobaan.csv")
df.head()
```

Kode diatas akan menghasilkan Output yaitu

	Temp	Yield	
0	50	3.3	
1	50	2.8	
2	50	2.9	
3	70	2.3	
4	70	2.6	

Gambar 39. Output Data Percobaan

, nama variabel Gambar 39 diganti dari **Temp** menjadi *x* dan **Yield** (Hasil Percobaan) menjadi *y*. Berikut adalah kode pemrogramannya

```
#Mengganti nama variabel
df.rename(columns={'Temp':'x','Yield':'y'}, inplace=True)
df.head()
```

Kode diatas akan menghasilkan output yaitu

	x	y
0	50	3.3
1	50	2.8
2	50	2.9
3	70	2.3
4	70	2.6

Gambar 40. Mengganti Nama Variabel di Data Hasil Percobaan

C. Membentuk Model Regresi Polynomial hingga Variabel Independen Signifikan terhadap Variabel dependen

Langkah selanjutnya memodelkan data set dengan regresi orde pertama atau dikenal sebagai regresi linier sederhana. Kode pemrograman python adlaah sebagai berikut.

```
#Memodelkan dengan Regresi Linier Berganda
import numpy as np
import statsmodels.api as sm
import statsmodels.formula.api as smf
x = df[['x']]
y = df['y']
reg_mod = 'y ~ x'
model = smf.ols(formula=reg_mod, data=df).fit()
print_model = model.summary()
print(print_model)
```

Hasil output dari Kode pemrograman Python diatas adalah

OLS Regression Results						
Dep. Variable:	y	R-squared:	0.092			
Model:	OLS	Adj. R-squared:	0.023			
Method:	Least Squares	F-statistic:	1.324			
Date:	Thu, 17 Nov 2022	Prob (F-statistic):	0.271			
Time:	19:26:35	Log-Likelihood:	-6.1371			
No. Observations:	15	AIC:	16.27			
Df Residuals:	13	BIC:	17.69			
Df Model:	1					
Covariance Type:	nonrobust					
coef	std err	t	P> t	[0.025	0.975]	
Intercept	2.3063	0.469	4.917	0.000	1.293	3.320
x	0.0068	0.006	1.151	0.271	-0.006	0.019
Omnibus:	0.182	Durbin-Watson:	1.267			
Prob(Omnibus):	0.913	Jarque-Bera (JB):	0.382			
Skew:	-0.111	Prob(JB):	0.826			
Kurtosis:	2.251	Cond. No.	371.			

Gambar 41. Output Model Regresi Polynomial orde pertama

Gambar 41 menunjukkan nilai p-value dari Uji F statistic sebesar 0.271 yang lebih besar dibandingkan $\alpha = 0.05$, oleh karena itu variabel independennya yaitu temperature belum signifikan terhadap hasil percobaan. Kemudian dilanjutkan dengan pemodelan polynomial orde ke-2 atau kuadratik. Kode pemrogramnya adalah

```
#Memodelkan dengan Regresi Linier Berganda
import numpy as np
import statsmodels.api as sm
import statsmodels.formula.api as smf
x = df[['x']]
y = df['y']
reg_mod = 'y ~ x+I(x**2)'
model = smf.ols(formula=reg_mod, data=df).fit()
print_model = model.summary()
print(print_model)
```

Hasil output dari kode diatas dapat dilihat pada Gambar 42.

OLS Regression Results						
Dep. Variable:	y	R-squared:	0.673			
Model:	OLS	Adj. R-squared:	0.619			
Method:	Least Squares	F-statistic:	12.36			
Date:	Thu, 17 Nov 2022	Prob (F-statistic):	0.00122			
Time:	19:34:22	Log-Likelihood:	1.5238			
No. Observations:	15	AIC:	2.952			
Df Residuals:	12	BIC:	5.077			
Df Model:	2					
Covariance Type:	nonrobust					
coef	std err	t	P> t	[0.025	0.975]	
Intercept	7.9605	1.259	6.323	0.000	5.218	10.703
x	-0.1537	0.035	-4.399	0.001	-0.230	-0.078
I(x ** 2)	0.0011	0.000	4.618	0.001	0.001	0.002
Omnibus:	0.618	Durbin-Watson:	2.248			
Prob(Omnibus):	0.734	Jarque-Bera (JB):	0.585			
Skew:	0.027	Prob(JB):	0.746			
Kurtosis:	2.034	Cond. No.	1.37e+05			

Gambar 42. Output Model Regresi Polynomial orde ke-dua

Gambar 42 menunjukkan nilai p-value dari Uji F statistic sebesar 0.001 yang lebih kecil dibandingkan $\alpha = 0.05$, oleh karena itu variabel independennya yaitu temperature telah signifikan terhadap hasil percobaan. Sehingga bisa kita katakan bahwa model terbaik untuk menyelidiki hubungan temperature dengan hasil percobaan adalah dengan model polynomial orde kedua. Berikut adalah model regresi terbaiknya

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 \quad (34)$$

$$\hat{y} = 7.9605 - 0.1537x + 0.0011x^2$$

Dari Persamaan (31) dan (32) kita akan dapat menghasilkan nilai prediksi dan residualnya. Kode pemrograman python untuk menghasilkan nilai prediksi \hat{y} adalah

```
prediksi = model.predict(x)
print(prediksi.head())
```

Output hasil kode pemrograman diatas antara lain

0	2.999686
1	2.999686
2	2.999686
3	2.336478
4	2.336478

```
dtype: float64
```

Adapun kode pemrograman python untuk menghasilkan nilai residual e_i adalah

```
residual=model.resid
print(residual.head())
```

Output hasil kode pemrograman diatas antara lain

```
0      0.300314
1     -0.199686
2     -0.099686
3     -0.036478
4      0.263522
dtype: float64
```

D. Perhitungan R-Squared

Pada Gambar 42, diketahui bahwa nilai R-squared sebesar 67.3. Artinya bahwa variabel hasil percobaan dapat dijelaskan oleh variabel temperatur sebesar 67.3%, sisanya 32.7% dijelaskan oleh variabel lainnya yang tidak diketahui.

E. Memberikan Hasil Kesimpulan

Tujuan penelitian menggunakan metode regresi variabel polinomial adalah ingin menyelidiki hubungan antara temperature dengan hasil percobaan. Berikut adalah model yang didapatkan:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$$

$$\hat{y} = 7.9605 - 0.1537x + 0.0011x^2$$

Berdasarkan nilai p-value dari Uji statistic F model diatas didapatkan nilai kurang dari 0.05 pada model kuadratiknya, sehingga dapat disimpulkan bahwa temperatur berhubungan secara kuadratik dengan hasil percobaan.

VI. Instruksi Tugas

Setelah memahami langkah-langkah penyelesaian permasalahan data menggunakan metode regresi Polinomial dengan bahasa pemrograman Python, pengguna e-modul praktikum dapat menyelesaikan tugas dengan instruksi sebagai berikut

- 1) Melakukan analisis data dengan bahasa pemrograman Python memanfaatkan layanan Google Colab
- 2) Hasil jawaban soal diketik menggunakan File Word ukuran A4 dengan format sebagai berikut:
 - a. Cover Laporan memuat Judul, Identitas pengguna, dan, Identitas Institusi
 - b. Daftar Isi, jika ada tabel dan gambar maka harus ada Daftar Tabel dan Daftar Gambar
 - c. Bab I Pendahuluan memuat latar belakang, tujuan, dan manfaat
 - d. Bab II Tinjauan Pustaka memuat teori dari metode
 - e. Bab III Metodologi data memuat dataset yang digunakan serta Langkah-langkah penyelesaian
 - f. Bab IV Hasil analisis memuat hasil penyelesaian berdasarkan soal permasalahan
 - g. Bab V Kesimpulan memuat hasil kesimpulan berdasarkan bab IV. Kesimpulan tidak boleh hasil dari copy paste
 - h. Bab VI Daftar Pustaka
- 3) Laporan dikumpulkan berbentuk Pdf.
- 4) File Notebook Google Colab hasil pemrograman Python juga ikut dikumpulkan dengan laporan
- 5) Laporan dan File Notebook Google Colab dimasukan kedalam satu folder dimana folder tersebut diberi nama identitas pengguna.

VI. Soal Regresi Variabel Polinomial

Mahasiswa Biologi mendapatkan tugas dari dosennya untuk meneliti umur dan Panjang 80 ikan secara random. Mereka ingin mengetahui apakah usia suatu ikan mempengaruhi ukuran panjangnya. Diketahui bahwa ternyata usia dan

Panjang ikan tidak berhubungan secara linier, sehingga pada data ini akan dicoba menganalisis menggunakan regresi polynomial. Berikut adalah data yang akan digunakan

Tabel 13. Data Ukuran Ikan Berdasarkan Usia

No	Usia	Panjang	No	Usia	Panjang
1	1	67	40	4	169
2	1	62	41	4	167
3	2	109	42	5	188
4	2	83	43	2	100
5	2	91	44	2	109
6	2	88	45	4	150
7	3	137	46	3	140
8	3	131	47	4	170
9	3	122	48	3	150
10	3	122	49	4	140
11	3	118	50	4	140
12	3	115	51	4	150
13	3	131	52	4	150
14	3	143	53	3	140
15	3	142	54	3	150
16	2	123	55	3	150
17	3	122	56	4	150
18	4	138	57	4	160
19	4	135	58	3	140
20	4	146	59	4	150
21	4	146	60	5	170
22	4	145	61	4	150
23	4	145	62	5	150
24	4	144	63	4	150
25	4	140	64	4	150
26	4	150	65	3	150
27	4	152	66	5	150
28	4	157	67	5	160
29	4	155	68	4	140
30	4	153	69	5	160
31	4	154	70	3	130
32	4	158	71	4	160
33	4	162	72	3	130
34	4	161	73	4	170
35	4	162	74	6	170
36	4	165	75	4	160
37	4	171	76	5	180
38	5	171	77	4	160
39	4	162	78	4	170

Kerjakan Data tersebut berdasarkan tujuan dari Mahasiswa dan dalam menganalisis, langkah-langkah harus diselesaikan secara lengkap sesuai dengan sub bab tutorial metode pada bab ini.

PENUTUP

Statistika regresi merupakan mata kuliah yang mempelajari konsep analisis regresi untuk menyelesaikan permasalahan yang berkaitan dengan hubungan antara dua variabel dependen (atau variabel respon) dengan variabel independen (atau variabel prediktor). Mata kuliah ini tidak hanya mempelajari teori regresi saja namun juga mempraktekan dengan data real menggunakan bahasa pemrograman Python. Oleh karenanya dibutuhkan suatu perangkat pembelajaran yang tidak hanya berisikan teori namun juga dapat melakukan praktikum menggunakan data dari berbagai bidang. Adapun perangkat pembelajaran yang saat ini dibutuhkan mahasiswa adalah berbentuk e-modul praktikum karena perkembangan jaman yang saat ini serba digital. Penyusunan e-modul praktikum ini diharapkan dapat membantu mahasiswa secara mandiri untuk mengelola, menganalisis, dan model regresi dari data atau informasi hasil pengamatan, serta dapat memprediksi dan mengetahui pengaruh variabel independennya terhadap variabel dependen.

Dalam penyusunan e-modul praktikum ini diharapakan dapat digunakan sebagai referensi tambahan dalam proses pembelajaran sehingga mahasiswa lebih mendalami materi perkuliahan dengan baik. Semoga e-modul praktikum ini bermanfaat bagi mahasiswa program studi Sains Data lebih mengembangkan diri untuk mencapai salah satu profil lulusan yang ditentukan yaitu sebagai data analyst. Penulis menyadari bahwa e-modul praktikum ini banyak kekurangan sehingga mohon saran dan kritik yang membangun demi sempurnanya penyusunan e-modul praktikum ini di masa yang akan datang.

KUNCI JAWABAN

Berikut adalah kunci dari sebagian jawaban soal yang telah diberikan pada BAB II hingga BAB VII antara lain:

BAB II : KORELASI

Koefisien Pearson: 0.96899

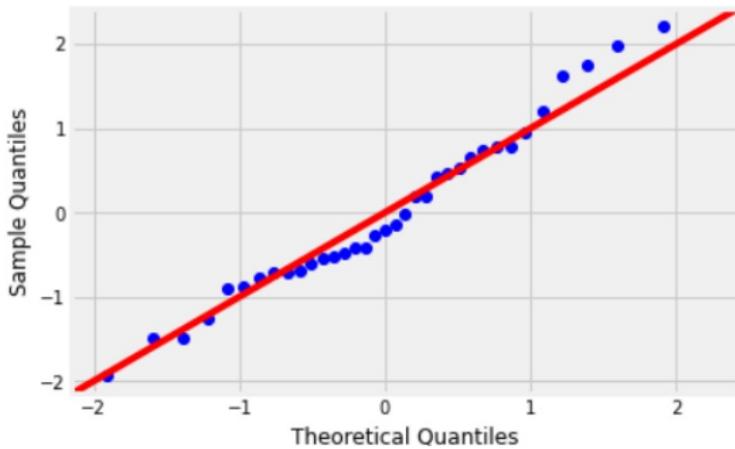
Koefisien Spearman: 0.98299

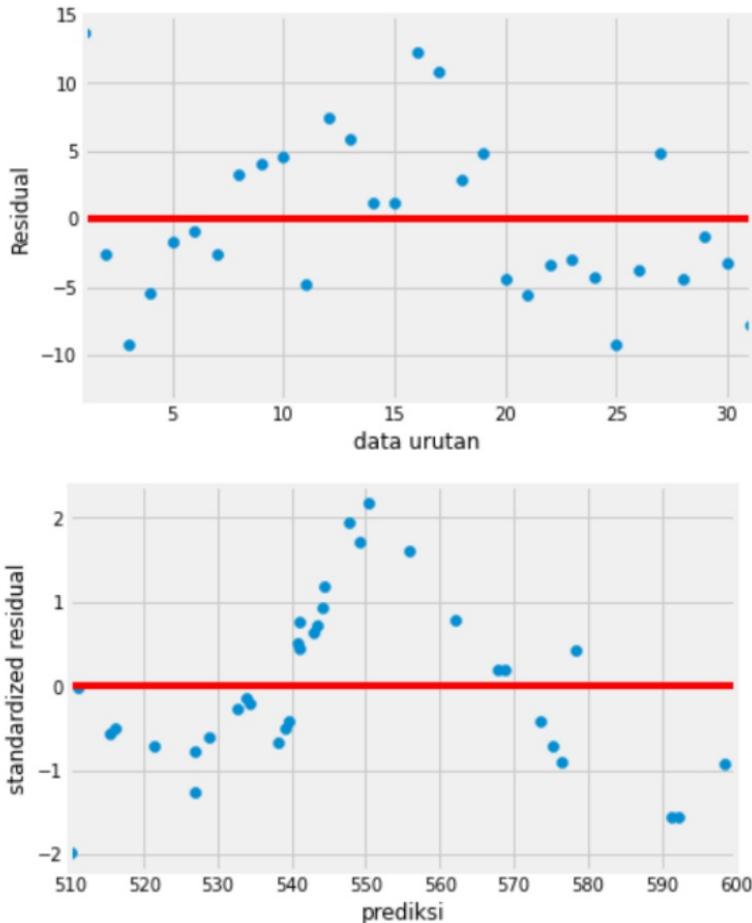
Koefisien Tau Kendall: 0.93990

BAB III : REGRESI LINIER SEDERHANA

OLS Regression Results						
Dep. Variable:	y	R-squared:	0.932			
Model:	OLS	Adj. R-squared:	0.930			
Method:	Least Squares	F-statistic:	455.0			
Date:	Fri, 18 Nov 2022	Prob (F-statistic):	7.08e-21			
Time:	09:01:40	Log-Likelihood:	-113.51			
No. Observations:	35	AIC:	231.0			
Df Residuals:	33	BIC:	234.1			
Df Model:	1					
Covariance Type:	nonrobust					
coef	std err	t	P> t	[0.025	0.975]	
const	479.7566	3.351	143.172	0.000	472.939	486.574
x	0.1201	0.006	21.331	0.000	0.109	0.132
Omnibus:	1.124	Durbin-Watson:	1.468			
Prob(Omnibus):	0.570	Jarque-Bera (JB):	1.108			
Skew:	0.377	Prob(JB):	0.575			
Kurtosis:	2.561	Cond. No.	1.85e+03			

BAB IV : ASUMSI RESIDUAL ANALISIS REGRESI LINIER





```
[('Jarque-Bera', 1.108423488848359),
 ('Chi^2 two-tail prob.', 0.5745249552531153),
 ('Skew', 0.3765341006018152),
 ('Kurtosis', 2.560739137572538)]
```

Durbin-Watson: 1.4684302824383628

```
[('Lagrange multiplier statistic', 0.2542097345383665),
 ('p-value', 0.6141263489657378),
 ('f-value', 0.24143705397614382),
 ('f p-value', 0.6264224165993728)]
```

BAB V : REGRESI LINIER BERGANDA

```
OLS Regression Results
=====
Dep. Variable:                  y   R-squared:          0.824
Model:                          OLS  Adj. R-squared:      0.815
Method: Least Squares          F-statistic:         86.84
Date: Sat, 19 Nov 2022          Prob (F-statistic): 1.06e-14
Time: 09:47:14                 Log-Likelihood:     55.811
No. Observations:               40   AIC:                -105.6
Df Residuals:                  37   BIC:                -100.6
Df Model:                      2
Covariance Type:               nonrobust
=====
            coef    std err        t      P>|t|      [ 0.025   0.975]
-----
const     -2.2435    0.263   -8.527      0.000    -2.777   -1.710
x1        0.0085    0.003    2.874      0.007     0.003    0.015
x2        0.0064    0.001    4.319      0.000     0.003    0.009
=====
Omnibus:                   0.282   Durbin-Watson:       1.749
Prob(Omnibus):              0.868   Jarque-Bera (JB):  0.030
Skew:                      0.066   Prob(JB):           0.985
Kurtosis:                   3.024   Cond. No.        8.98e+03
=====
```

BAB VI: REGRESI DUMMY

```
OLS Regression Results
=====
Dep. Variable:                  y   R-squared:          0.891
Model:                          OLS  Adj. R-squared:      0.885
Method: Least Squares          F-statistic:         151.2
Date: Sat, 19 Nov 2022          Prob (F-statistic): 1.57e-18
Time: 10:06:46                 Log-Likelihood:     -374.87
No. Observations:               40   AIC:                755.7
Df Residuals:                  37   BIC:                760.8
Df Model:                      2
Covariance Type:               nonrobust
=====
            coef    std err        t      P>|t|      [ 0.025   0.975]
-----
Intercept  -1.027e+04  1870.696   -5.489      0.000   -1.41e+04   -6477.928
x1         182.3231   18.226    10.004      0.000    145.394    219.252
x2_1       4233.8511  1310.135    3.232      0.003    1579.266    6888.436
=====
Omnibus:                   8.791   Durbin-Watson:       1.191
Prob(Omnibus):              0.012   Jarque-Bera (JB):  8.392
Skew:                      0.789   Prob(JB):           0.0151
Kurtosis:                   4.595   Cond. No.        540.
=====
```

BAB VII: REGRESI POLYNOMIAL

```
OLS Regression Results
=====
Dep. Variable:                  y   R-squared:                 0.735
Model:                          OLS   Adj. R-squared:            0.731
Method: Least Squares          F-statistic:                210.7
Date: Sat, 19 Nov 2022          Prob (F-statistic):        1.31e-23
Time: 10:35:45                 Log-Likelihood:             -306.73
No. Observations:               78    AIC:                      617.5
Df Residuals:                  76    BIC:                      622.2
Df Model:                      1
Covariance Type:               nonrobust
=====
            coef      std err      t      P>|t|      [0.025      0.975]
-----
Intercept  62.6490     5.755     10.887     0.000     51.188     74.110
x          22.3123    1.537     14.514     0.000     19.251     25.374
=====
Omnibus:                   4.087   Durbin-Watson:           1.243
Prob(Omnibus):              0.130   Jarque-Bera (JB):       2.372
Skew:                     -0.191   Prob(JB):                 0.305
Kurtosis:                   2.236   Cond. No.                  16.2
=====
```

DAFTAR PUSTAKA

- Haslwanter, Thomas. 2016. *An Introduction to Statistics with Python : With Application in the Life Sciences*. Switzerland: Springer International Publishing Switzerland
- Kutner, M., Nachtsheim, C.J., dan Neter, J. 2004. *Applied Linier Regression Models*. Edisi Keempat. New York: Mc Graw- Hill/Irwin.
- Massaron, Luca, dan B, Alberto. 2016. *Regression Analysis with Python*. Brimingham: Packt Publishing
- Montgomery, D.C., Peck, E.A., Vining, G.G. 2012. *Introduction to Linear Regression Analysis*. Edisi Kelima. New Jersey: John Wiley & Sons, Inc.
- Weiers, Ronald M. 2008. *Introduction to Business Statistics*. USA: South-Western Cengage Learning
- Yan, X., dan Su, X.G. 2009. *Linear Regression Analysis: Theory and Computing*. Singapura: World Scientific Publishing Co. Pte. Ltd.
- Zaid, Mohamed Ahmed. 2015. *Correlation and Regression Analysis*. Turkey: Organization of Islamic Cooperation