

Instruksi:

Carilah data dengan permasalahan yang dapat dimodelkan dengan regresi linear. Data memuat minimal 200 pengamatan, dengan pengukuran numerik maupun kategorik. Dapat menggunakan data pada Project 1. Prosedur yang dilakukan mencakup pemilihan model terbaik (dengan metrik yang sesuai), seleksi variabel, pengecekan asumsi (dan langkah mengatasi masalah jika ada pelanggaran asumsi), pastikan bahwa tidak ada *pitfalls* pada regresi yang diajukan.

Lakukan pengolahan data (jika diperlukan lakukan pre-processing data terlebih dahulu), dengan prosedur yang tepat, kemudian lakukan analisis dan interpretasi hasilnya. Ikuti langkah – langkah berikut.

Bagian 1. Pendahuluan

[Tuliskan dengan baik masalah apa yang akan dibahas, sumber data (*berikan link sumber data, atau lampirkan di Bagian 6*), ukuran data, dan jumlah pengukuran (kolom data), skala/tipe data, dan arti/maksud dari pengukuran-pengukuran tersebut (jika diketahui). Tentukan variabel respon dan variabel prediktor. Tuliskan tujuan dari analisis regresi yang akan dilakukan.]

Bagian 2. Pre-processing (jika ada) dan analisis deskriptif

[Tuliskan langkah-langkah apa yang dilakukan untuk pre-processing data, kenapa (atau untuk apa) langkah tersebut dilakukan, hasil apa yang diperoleh dari langkah tersebut. Tuliskan juga keterkaitan antara 1 proses dengan proses lainnya (jika terkait). Tuliskan permasalahan apa yang ada saat pre-processing data. Lakukan eksplorasi dan visualisasi yang diperlukan; berdasarkan hasil ini tuliskan hipotesis yang akan dicek dalam melakukan regresi pada proses selanjutnya. (Codes R/Python dilampirkan)]

Bagian 3. Pemodelan

[Tuliskan model regresi (bisa 1 atau lebih) yang diajukan (lengkap dengan penjelasan notasi dan asumsi), dan alasan mengajukan model tersebut.

Jelaskan alasan jika perlu dilakukan transformasi variabel, dan dasar pemilihan fungsi untuk transformasi variabel.]

Bagian 4. Pengolahan data dan analisis hasil

- Tuliskan analisis apa saja yang dilakukan, kenapa/hasil apa yang diharapkan (dan yang diperoleh) dari melakukan analisis tersebut (Codes R/Python dilampirkan).
- Apakah hasil analisis mendukung hipotesis pada **Bagian 2**? Informasi bermakna apa saja yang dihasilkan dari proses ini?
- Apakah model yang diajukan sudah merupakan model yang terbaik? Jelaskan prosedur yang dilakukan untuk meyakinkan bahwa model yang diperoleh adalah model terbaik. Jelaskan dasar penentuan kriteria terbaik.
- Apakah ada asumsi yang tidak terpenuhi? Jika ada, apa tindak lanjut yang dilakukan untuk mengatasi hal tersebut?
- Tuliskan insight/informasi yang berguna yang diperoleh dari regresi yang diperoleh.

Bagian 5. Penutup

[Tuliskan kesimpulan apa yang kalian dapatkan dari penugasan ini, kaitkan dengan permasalahan yang diajukan pada **Bagian 1**.]

Bagian 6. Lampiran

Berikan link GDrive folder yang berisi file data, codes, dan file presentasi. Set-up Gdrive: anyone with the link can assess, sehingga dosen tidak perlu ask permission untuk akses folder tersebut.

Penilaian Presentasi:

- Dilakukan pada sesi kelas masing-masing di Minggu kedua periode UAS.
- Durasi waktu Max 7 menit.
- Komponen penilaian mengacu pada rubrik yang ada (lihat file excel: Rubrik penilaian presentasi final)
- Urutan presentasi kelompok dilakukan secara acak

LAPORAN PROJECT 2

**ANALISIS DAN MODEL REGRESI PREDIKSI CPU
PERFORMANCE**



Dosen mata kuliah : Ibu Sarini Abdullah

Disusun oleh :

Jason Justin Andryana	(2206029670)
Bryan Reynaldy	(2206029613)
Rachelle Melody d'Lyra Soentara	(2206051456)
Soraya Indira Putri Djabbar	(2206053902)
Yohanes Nathael	(2206051405)

Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Indonesia
2023

Anggota kelompok:

No	Nama	NPM	Kontribusi	Tingkat kontribusi
1	Bryan Reynaldy	2206029613	Coding (modelling), Bab 1, Bab 2 dan penutup	100%
2	Jason Justin	2206029670	Coding (compare model)	100%
3	Rachelle Melody	2206051456	Mencari data, <i>coding</i> (modelling, preprocessing, uji asumsi), membuat Bab 4	100%
4	Soraya Indira	2206053902	Membuat Bab 1 dan penutup	100%
5	Yohanes Nathael	2206051405	Coding (modelling, validation), Bab 3	100%

DAFTAR ISI

BAB 1	7
1.1 Latar Belakang	7
1.2 Rumusan Masalah	7
1.3 Data dan Variabel	7
BAB 2	9
2.1 Statistika Deskriptif	9
2.2 Preprocessing Data	9
2.3 Deskripsi Data	11
2.4 Visualisasi Data	11
2.4.1 Corellation Matrix	11
2.4.2 Pair plot	12
2.5 Hipotesis	12
BAB 3	13
3.1 Identifikasi Multikolinearitas	13
3.2 Variable Selection	14
3.2.1 Backward Regression	14
3.2.2 Forward Regression	14
3.3 Model 1 : General Form = Forward Regression	15
3.4 Model 2 : Transformation Respon Model	16
3.5 Model 3 : Polynomial Model	16
3.6 Model 4 : Backward Regression	17
BAB 4	19
4.1 Uji Asumsi Model 1	19
4.1.1 Mean of Residual	19
4.1.2 Homoscedasticity	19
Uji Goldfeld Quandt	19
4.1.3 Normality of Error Terms	20
4.2 Uji Asumsi Model 2	Error! Bookmark not defined.
4.2.1 Mean of Residual	Error! Bookmark not defined.
4.2.2 Homoscedasticity	Error! Bookmark not defined.
Uji Goldfeld Quandt	Error! Bookmark not defined.
4.2.3 Normality of Error Terms	Error! Bookmark not defined.
4.3 Uji Asumsi Model 3	20
4.3.1 Mean of Residual	21
4.3.2 Homoscedasticity	Error! Bookmark not defined.
4.3.3 Normality of error terms	21
4.4 Uji Asumsi Model 4	22

4.4.1 Mean of Residual	22
4.4.2 Homoscedasticity	22
4.4.3 Normality of error terms	23
BAB 5	26
Penutup.....	26
BAB 6	30
Lampiran	30

BAB 1

Pendahuluan

1.1 Latar Belakang

Perkembangan teknologi komputer dan peningkatan dalam CPU telah menjadi pendorong utama dalam peningkatan kinerja sistem komputasi. Pemahaman mendalam tentang faktor-faktor yang memengaruhi kinerja CPU sangat penting dalam pengembangan sistem komputer yang efisien dan andal. Waktu siklus, ukuran memori, dan saluran memori (memory channels) adalah beberapa faktor kunci yang dikenal memengaruhi kinerja CPU.

Waktu siklus mencerminkan seberapa cepat CPU dapat mengeksekusi instruksi dasar, ukuran memori mempengaruhi kemampuan sistem untuk menangani beban kerja yang intensif, dan saluran memori memainkan peran penting dalam transfer data antara CPU dan memori. Penelitian sebelumnya telah menunjukkan bahwa hubungan kompleks antara variabel-variabel ini dan kinerja CPU dapat dijelaskan melalui pendekatan analisis regresi.

Oleh karena itu, penelitian ini bertujuan untuk memodelkan kinerja CPU berdasarkan waktu siklus, ukuran memori, saluran, dan faktor-faktor lainnya menggunakan teknik regresi.


1.2 Rumusan Masalah

1. Faktor-faktor apa saja yang mempengaruhi kinerja CPU?
2. Bagaimana hubungan antara variable factor-faktor tersebut dalam meningkatkan performa CPU?

1.3 Data dan Variabel

Beberapa informasi terkait data dan variabel yang kami gunakan dalam penelitian ini adalah sebagai berikut:

1. Dataset dari penelitian kami diperoleh melalui UC Irvine Machine Learning Repository dengan tautan : <https://archive.ics.uci.edu/dataset/29/computer+hardware>
2. Baris dan kolom dari dataset berjumlah 209 baris dan 10 kolom.



	VendorName	ModelName	MYCT	MMIN	MMAx	CACH	CHMIN	CHMAX	PRP	ERP
0	adviser	32/60	125	256	6000	256	16	128	198	199
1	amdahl	470v/7	29	8000	32000	32	8	32	269	253
2	amdahl	470v/7a	29	8000	32000	32	8	32	220	253
3	amdahl	470v/7b	29	8000	32000	32	8	32	172	253
4	amdahl	470v/7c	29	8000	16000	32	8	16	132	132
...
204	sperry	80/8	124	1000	8000	0	1	8	42	37
205	sperry	90/80-model-3	98	1000	8000	32	2	8	46	50
206	sratus	32	125	2000	8000	0	2	14	52	41
207	wang	vs-100	480	512	8000	32	0	0	67	47
208	wang	vs-90	480	1000	4000	0	0	0	45	25

209 rows × 10 columns

3. Definisi variabel adalah sebagai berikut:

Nama Variable	Peran	Tipe	Deskripsi	Unit
VendorName	Feature	Kategorik	Adviser, amdahl,apollo, basf, bti, burroughs, c.r.d, cambex, cdc, dec, dg, formation, four-phase, gould, honeywell, hp, ibm, ipl, magnuson, microdata, nas, ncr, nixdorf, perkin-elmer, prime, siemens, sperry, sratus, wang	-
ModelName	Feature	Kategorik	many unique symbols	-
MYCT	Feature	Integer	machine cycle time	nanoseconds
MMIN	Feature	Integer	minimum main memory	kilobytes
MMAX	Feature	Integer	maximum main memory	kilobytes
CACH	Feature	Integer	cache memory	kilobytes
CHMIN	Feature	Integer	minimum channels	units
CHMAX	Feature	Integer	maximum channels	units
PRP	Feature	Integer	published relative performance	-
ERP	Feature	Integer	estimated relative performance from the original article	-

4. Variabel respon atau variabel dependen dari penelitian ini adalah PRP (Published Relative Performance)
5. Variabel prediktor atau variabel independen dari penelitian ini adalah:
 - MYCT (Machine Cycle Time)
 - MMIN (Minimum Main Memory)
 - MMAX (Maximum Main Memory)
 - CACH (Cache Memory)
 - CHMIN (Minimum Channels)
 - CHMAX (Maximum Channels)

BAB 2

Pre-Processing dan Statistika Deskriptif

2.1 Statistika Deskriptif

Pre-processing data regresi linear berganda merupakan tahap yang harus dilakukan sebelum melaksanakan analisis regresi. *Pre-processing* melibatkan beberapa tahapan seperti memeriksa *missing values*, tipe data, memeriksa data yang terduplikasi, visualisasi data, dan lain-lain. Tahapan ini berguna untuk menyiapkan data agar lebih mudah dan lebih siap untuk dianalisis.

```
[ ] df.describe()
```

	MYCT	MMIN	MMAx	CACH	CHMIN	CHMAX	PRP
count	209.000000	209.000000	209.000000	209.000000	209.000000	209.000000	209.000000
mean	203.822967	2867.980861	11796.153110	25.205742	4.698565	18.267943	105.622010
std	260.262926	3878.742758	11726.564377	40.628722	6.816274	25.997318	160.830733
min	17.000000	64.000000	64.000000	0.000000	0.000000	0.000000	6.000000
25%	50.000000	768.000000	4000.000000	0.000000	1.000000	5.000000	27.000000
50%	110.000000	2000.000000	8000.000000	8.000000	2.000000	8.000000	50.000000
75%	225.000000	4000.000000	16000.000000	32.000000	6.000000	24.000000	113.000000
max	1500.000000	32000.000000	64000.000000	256.000000	52.000000	176.000000	1150.000000

2.2 Preprocessing Data

Terdapat beberapa langkah yang kami lakukan untuk *pre-processing* data sebagai berikut :

1. Memeriksa tipe data

```
6] df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 209 entries, 0 to 208
Data columns (total 10 columns):
#   Column      Non-Null Count  Dtype
---  ---
0   VendorName  209 non-null   object
1   ModelName   209 non-null   object
2   MYCT        209 non-null   int64
3   MMIN        209 non-null   int64
4   MMAx        209 non-null   int64
5   CACH        209 non-null   int64
6   CHMIN       209 non-null   int64
7   CHMAX       209 non-null   int64
8   PRP         209 non-null   int64
9   ERP         209 non-null   int64
dtypes: int64(8), object(2)
memory usage: 16.5+ KB
```

2. Memeriksa kolom yang relevan

Akan diperiksa jumlah perbedaan dalam kolom VendorName dan ModelName

```
| df['VendorName'].unique()
```

```
array(['adviser', 'amdahl', 'apollo', 'basf', 'bti', 'burroughs', 'c.r.d',
      'cdc', 'cambex', 'dec', 'dg', 'formation', 'four-phase', 'gould',
      'hp', 'harris', 'honeywell', 'ibm', 'ipl', 'magnuson', 'microdata',
      'nas', 'ncr', 'nixdorf', 'perkin-elmer', 'prime', 'siemens',
      'sperry', 'sratus', 'wang'], dtype=object)
```

```
df['ModelName'].unique()

array(['32/60', '470v/7', '470v/7a', '470v/7b', '470v/7c', '470v/b',
      '580-5840', '580-5850', '580-5860', '580-5880', 'dn320', 'dn420',
      '7/65', '7/68', '5000', '8000', 'b1955', 'b2900', 'b2925', 'b4955',
      'b5900', 'b5920', 'b6900', 'b6925', '68/10-80', 'universe:2203t',
      'universe:68', 'universe:68/05', 'universe:68/137',
      'universe:68/37', 'cyber:170/750', 'cyber:170/760',
      'cyber:170/815', 'cyber:170/825', 'cyber:170/835', 'cyber:170/845',
      'omega:480-i', 'omega:480-ii', 'omega:480-iii', '1636-1',
      '1636-10', '1641-1', '1641-11', '1651-1', 'decsys:10:1091',
      'decsys:20:2060', 'microvax-1', 'vax:11/730', 'vax:11/750',
      'vax:11/780', 'eclipse:c/350', 'eclipse:m/600', 'eclipse:mv/10000',
      'eclipse:mv/4000', 'eclipse:mv/6000', 'eclipse:mv/8000',
      'eclipse:mv/8000-ii', 'f4000/100', 'f4000/200', 'f4000/200ap',
      'f4000/300', 'f4000/300ap', '2000/260', 'concept:32/8705',
      'concept:32/8750', 'concept:32/8780', '3000/30', '3000/40',
      '3000/44', '3000/48', '3000/64', '3000/88', '3000/iii', '100',
      '300', '500', '600', '700', '80', '800', 'dps:6/35', 'dps:6/92',
      'dps:6/96', 'dps:7/35', 'dps:7/45', 'dps:7/55', 'dps:7/65',
      'dps:8/44', 'dps:8/49', 'dps:8/50', 'dps:8/52', 'dps:8/62',
      'dps:8/20', '3033:s', '3033:u', '3081', '3081:d', '3083:b',
      '3083:e', '370/125-2', '370/148', '370/158-3', '38/3', '38/4',
```

Karena banyaknya macam perbedaan dalam kedua kolom ini, maka dapat diketahui bahwa kedua kolom ini hanyalah informasi tambahan. Kolom ini akan didrop karena tidak berpengaruh dan bukan menjadi acuan dalam membuat prediksi model komputer terbaik. ERP adalah hasil prediksi model oleh peneliti sebelumnya, sehingga tidak akan digunakan dalam project ini. Oleh karena itu, kolom ERP akan didrop dari df.

3. Memeriksa *missing value*

Terlihat bahwa tidak ada *missing value* pada data

```
df.isnull().sum()

MYCT      0
MMIN      0
MMAX      0
CACH      0
CHMIN     0
CHMAX     0
PRP       0
dtype: int64
```

4. Memeriksa *duplicated data*

Terlihat bahwa tidak ada data yang terduplikasi

```
df.duplicated().sum()

0
```

2.3 Deskripsi Data

```
1 df.describe()
```

	MYCT	MMIN	MMAx	CACH	CHMIN	CHMAX	PRP
count	209.000000	209.000000	209.000000	209.000000	209.000000	209.000000	209.000000
mean	203.822967	2867.980861	11796.153110	25.205742	4.698565	18.267943	105.622010
std	260.262926	3878.742758	11726.564377	40.628722	6.816274	25.997318	160.830733
min	17.000000	64.000000	64.000000	0.000000	0.000000	0.000000	6.000000
25%	50.000000	768.000000	4000.000000	0.000000	1.000000	5.000000	27.000000
50%	110.000000	2000.000000	8000.000000	8.000000	2.000000	8.000000	50.000000
75%	225.000000	4000.000000	16000.000000	32.000000	6.000000	24.000000	113.000000
max	1500.000000	32000.000000	64000.000000	256.000000	52.000000	176.000000	1150.000000

Variable Insight :

- Melihat pengukuran masing-masing variabel, terlihat bahwa dalam kolom *CACH*, *CHMIN*, dan *CHMAX*, terdapat nilai = 0, yang akan diperiksa.
- Karena *CACH* adalah Cache Memory, maka nilai 0 masuk akal sehingga tidak diproses lebih lanjut.
- Karena *CHMIN* adalah minimal Channel, maka nilai 0 juga tidak bermasalah.

Akan diperiksa *CHMAX* yaitu maximal Channel yang bernilai 0

```
1 # Filter rows where "CACH" is equal to 0
2 filtered_df = df[df['CHMAX'] == 0]
3
4 # Display the filtered DataFrame in a markdown table format
5 display(filtered_df)
```

	MYCT	MMIN	MMAx	CACH	CHMIN	CHMAX	PRP
122	1500	768	1000	0	0	0	12
123	1500	768	2000	0	0	0	18
124	800	768	2000	0	0	0	20
207	480	512	8000	32	0	0	67
208	480	1000	4000	0	0	0	45

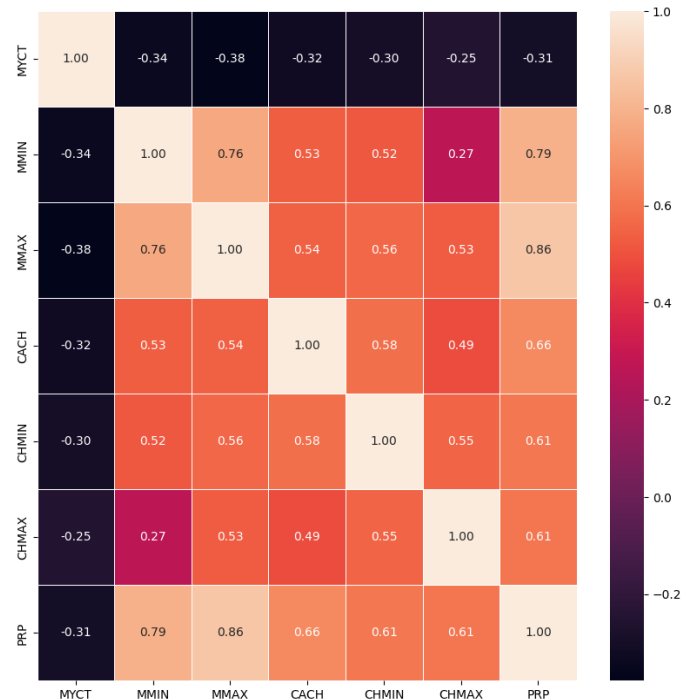
Karena pada *CHMAX* (Channel maksimum) yang bernilai 0, *CHMIN* (Channel minimum) juga 0, maka data tidak bermasalah.

2.4 Visualisasi Data

2.4.1 Correlation Matrix

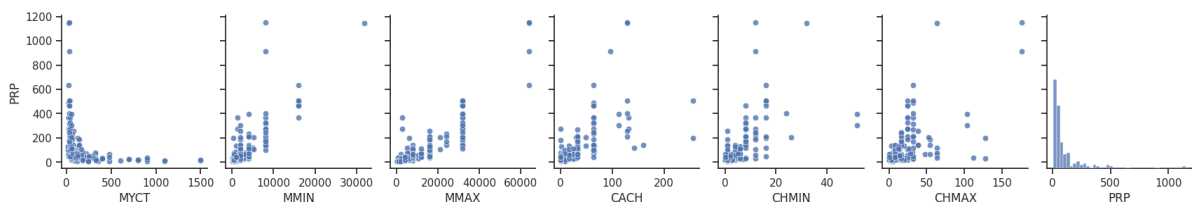
Dalam analisis korelasi matrix, ditemukan bahwa terdapat korelasi linear yang cukup baik antara variabel kinerja mesin (PRP) dengan sebagian besar prediktor, kecuali pada variabel MYCT. Oleh karena itu, langkah selanjutnya adalah melakukan pemeriksaan melalui visualisasi plot antara PRP

dan masing-masing prediktor guna mendapatkan wawasan lebih lanjut terkait hubungan yang teramati.



2.4.2 Pair plot

Melalui visualisasi pair plot, terdapat beberapa hubungan yang dapat terbentuk antara variabel kinerja mesin (PRP) dengan prediktor lainnya. Oleh karena itu, pair plot memberikan gambaran visual yang dapat digunakan untuk lebih memahami pola hubungan antara variable dependen dan independen dalam dataset ini.



2.5 Hipotesis

Dari visualisasi yang dilakukan, hipotesis yang akan diperiksa dalam melakukan regresi adalah

1. Asumsi dari model linear regresi terpenuhi
2. Model *general* additive kurang akurat dibandingkan model dengan polinom
3. Hubungan PRP dan MYCT adalah $\frac{1}{x}$, hubungan PRP dan MMAX adalah x^2

BAB 3

Pemodelan

3.1 Identifikasi Multikolinearitas

Sebelum dilakukan pemodelan, akan dilakukan pengecekan multikolinearitas dengan parameter *Variable Inflation Factor* (VIF). VIF digunakan untuk mengidentifikasi multikolinearitas dalam model regresi. Jika skor VIF dari predictor dalam model > 10 , maka dinyatakan terjadi multikolinearitas.

```
1 from statsmodels.stats.outliers_influence import variance_inflation_factor
2
3 # VIF dataframe
4 vif_data = pd.DataFrame()
5 vif_data["feature"] = features.columns
6
7 # calculating VIF for each feature
8 vif_data["VIF"] = [variance_inflation_factor(features.values, i)
9                    for i in range(len(features.columns))]
10
11 print(vif_data)
```

	feature	VIF
0	MYCT	1.067222
1	MMIN	4.493385
2	MMAx	5.890977
3	CACH	2.563899
4	CHMIN	2.890250
5	CHMAX	2.808009

Karena semua nilai VIF dari predictor-prediktor < 10 , maka multikolinearitas dianggap rendah (dapat diterima). Kesimpulannya adalah tidak ada indikasi kuat bahwa predictor-prediktor tersebut berkorelasi signifikan satu sama lain (Halaman 364 Mendenhall).

3.2 Variable Selection

Akan dilakukan prosedur variable selection yang bertujuan untuk memutuskan predictor-prediktor yang akan dimasukkan dalam model regresi multiple untuk memprediksi performa CPU. Dengan variable screening ini, dapat ditentukan variable independent mana yang paling penting untuk y dan yang kurang penting.

3.2.1 Backward Selection

Dalam Backward Regression, pada tiap iterasinya, akan dihilangkan variable yang tidak signifikan, yaitu variable dengan $p - value > \alpha = 0.05$ atau yang memiliki nilai t terkecil.

```
1 reg_back = backward(features, target)
2 print(f"Hasil stepwise regression backward: {reg_back}")
3
```

Iterasi 1

```
p-value CACH: 7.585892904337522e-06
p-value CHMAX: 1.646289357691269e-10
p-value CHMIN: 0.7523591684805726
p-value MMAX: 1.3178020394293263e-15
p-value MMIN: 9.419324701558424e-15
p-value MYCT: 0.005798460121219481
```

Iterasi 2

```
p-value CACH: 5.069767710313765e-06
p-value CHMAX: 3.060592052703342e-11
p-value MMAX: 1.17574301832507e-15
p-value MMIN: 4.343468608307895e-15
p-value MYCT: 0.005395138842076806
```

Hasil stepwise regression backward: ['CACH', 'CHMAX', 'MMAX', 'MMIN', 'MYCT']

Dapat dilihat bahwa dalam iterasi-1, $p - value 'CHMIN' = 0.752 > \alpha = 0.05$. Maka, akan dihilangkan variable ' $CHMIN$ ' dari model. Iterasi-2, $p - value$ pada kelima predictor sudah signifikan, yaitu $< \alpha = 0.05$. Oleh karena itu, iterasi dihentikan, dan menghasilkan model **reg_back**.

$$E(PRP) = \beta_0 + \beta_1 CACH + \beta_2 CHMAX + \beta_3 MMAX + \beta_4 MMIN + \beta_5 MYCT$$

3.2.2 Forward Selection

Dalam Forward Regression, pada tiap iterasinya, akan dimasukkan variable yang paling signifikan, yaitu variable dengan $p - value < \alpha = 0.05$ atau yang memiliki nilai t terbesar.

```
1 reg_forw = forward(features, target)
2 print(f"Hasil stepwise regression forward: {reg_forw}")
```

```
p-value CACH: 8.505256503661048e-28
p-value CHMAX: 7.876170373575824e-11
p-value CHMIN: 0.0003574001688335355
p-value MMAX: 6.282711642553545e-39
p-value MMIN: 4.6780254740220993e-14
p-value MYCT: 0.005798460121219481
```

Hasil stepwise regression forward: ['CACH', 'CHMAX', 'CHMIN', 'MMAX', 'MMIN', 'MYCT']

Dapat dilihat bahwa pada iterasi terakhir, terbentuk model dengan semua $p - value$ prediktornya signifikan yaitu $p - value$ pada keenam predictor sudah signifikan, yaitu semua nilainya $< \alpha = 0.05$. Oleh karena itu, iterasi dihentikan dan menghasilkan model **reg_forw**.

$$E(PRP) = \beta_0 + \beta_1 CACH + \beta_2 CHMAX + \beta_3 CHMIN + \beta_4 MMAX + \beta_5 MMIN + \beta_6 MYCT$$

3.3 Model 1 : General Form

Model pertama yang diajukan adalah model bentuk umum dari *multiple linear regression*, yang mencakup semua prediktor yang tersedia dalam dataset. Tujuan utama dari model ini adalah untuk mengidentifikasi signifikansi atau pengaruh dari masing-masing parameter (koefisien regresi) terhadap variabel dependen. Model ini sekaligus merupakan model yang terbentuk dari *variable selection metode forward*. Model ini memberikan pandangan awal tentang sejauh mana masing-masing prediktor berkontribusi dalam menjelaskan variasi dalam variabel dependen.

$$Y = \beta_0 + \beta_1 MYCT + \beta_2 MMIN + \beta_3 MMAX + \beta_4 CACH + \beta_5 CHIN + \beta_6 CHMAX + \epsilon ; \epsilon \sim NIID(0, \sigma^2)$$

- β_0 adalah intercept (konstanta) dari model
- $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6$ adalah koefisien regresi untuk masing-masing variabel independen "MYCT", "MMIN", "MMAX", "CACH", "CHMIN", dan "CHMAX".

$$\hat{y} = -55.8939 + 0.0489MYCT + 0.0153MMIN + 0.0056MMAX + 0.6414CACH - 0.2704CHMIN + 1.4825CHMAX$$

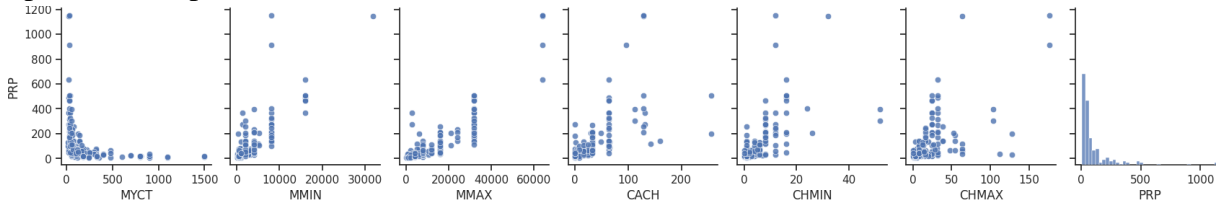
OLS Regression Results						
=====						
Dep. Variable:	PRP	R-squared:	0.865			
Model:	OLS	Adj. R-squared:	0.861			
Method:	Least Squares	F-statistic:	215.5			
Date:	Sat, 16 Dec 2023	Prob (F-statistic):	6.24e-85			
Time:	10:59:58	Log-Likelihood:	-1148.7			
No. Observations:	209	AIC:	2311.			
Df Residuals:	202	BIC:	2335.			
Df Model:	6					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	-55.8939	8.045	-6.948	0.000	-71.757	-40.031
MYCT	0.0489	0.018	2.789	0.006	0.014	0.083
MMIN	0.0153	0.002	8.371	0.000	0.012	0.019
MMAX	0.0056	0.001	8.681	0.000	0.004	0.007
CACH	0.6414	0.140	4.596	0.000	0.366	0.917
CHMIN	-0.2704	0.856	-0.316	0.752	-1.958	1.417
CHMAX	1.4825	0.220	6.737	0.000	1.049	1.916
=====						
Omnibus:	99.727	Durbin-Watson:	1.202			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1140.969			
Skew:	1.495	Prob(JB):	1.74e-248			
Kurtosis:	14.049	Cond. No.	3.32e+04			

Kami memilih model 1 dengan *fitting general form* dari model regresi dengan semua variabel untuk melihat signifikansi dari tiap parameter. Berdasarkan model ini, kami baru akan memodifikasi model selanjutnya.

3.4 Model 2 : Polynomial Model

Untuk memilih model selanjutnya, akan dilihat bagaimana hubungan dari tiap prediktor dengan respon. Transformasi akan bergantung pada bentuk plot data dan hubungan yang muncul pada plot prediktor-respon.



Dapat dilihat dari plot, bahwa hubungan dari PRP dengan prediktor MYCT similar dengan grafik $y = \frac{1}{x}$. Hal ini juga didukung dengan pengukuran bahwa $\frac{1}{MYCT} = \text{Clock Speed CPU}$. Juga dapat dilihat bahwa relasi antara PRP dengan prediktor MMAX membentuk *curvature*, maka akan diasumsikan PRP dan MMAX mendekati grafik $y = x^2$. Sedangkan, hubungan antara variable-variable predictor lainnya dengan PRP terlihat linear secara visual *scatter plot*. Maka, dibuatlah model dengan menggunakan *higher order polynomial* dari prediktor MMAX. Dan dibuat juga hubungan $\frac{1}{x}$ untuk variable MYCT.

$$Y = \beta_0 + \beta_1 MYCT + \beta_2 MMIN + \beta_3 MMAX + \beta_4 CACH + \beta_5 CHIN + \beta_6 CHMAX + \beta_7 \left(\frac{1}{MYCT} \right) + \beta_8 MMAX^2 + \epsilon ; \epsilon \sim NIID(0, \sigma^2)$$

- β_0 adalah intercept (konstanta) dari model
- $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6$ adalah koefisien regresi untuk masing-masing variabel independen "MYCT", "MMIN", "MMAX", "CACH", "CHMIN", dan "CHMAX".
- β_7 adalah koefisien regresi untuk $\frac{1}{MYCT}$, dan β_8 adalah koefisien regresi untuk $MMAX^2$.

$$\hat{y} = 8.2244 + 0.0098MYCT + 0.01MMIN - 0.0029MMAX + 0.8423CACH + 0.6509CHMIN + 0.6141CHMAX + 953.1275 \left(\frac{1}{MYCT} \right) + 0.000000188(MMAX^2)$$

OLS Regression Results						
=====						
Dep. Variable:	PRP		R-squared:	0.928		
Model:	OLS		Adj. R-squared:	0.926		
Method:	Least Squares		F-statistic:	324.4		
Date:	Sat, 16 Dec 2023		Prob (F-statistic):	4.12e-110		
Time:	12:45:10		Log-Likelihood:	-1082.3		
No. Observations:	209		AIC:	2183.		
Df Residuals:	200		BIC:	2213.		
Df Model:	8					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	8.2244	8.996	0.914	0.362	-9.514	25.963
MYCT	0.0098	0.015	0.661	0.509	-0.019	0.039
MMIN	0.0100	0.001	6.900	0.000	0.007	0.013
MMAX	-0.0029	0.001	-3.612	0.000	-0.004	-0.001
CACH	0.8423	0.104	8.121	0.000	0.638	1.047
CHMIN	0.6509	0.637	1.021	0.308	-0.606	1.908
CHMAX	0.6141	0.178	3.449	0.001	0.263	0.965
1/MYCT	953.1275	421.493	2.261	0.025	121.987	1784.268
MMAX^2	1.882e-07	1.42e-08	13.286	0.000	1.6e-07	2.16e-07
=====						
Omnibus:	47.513		Durbin-Watson:	1.611		
Prob(Omnibus):	0.000		Jarque-Bera (JB):	754.908		
Skew:	-0.120		Prob(JB):	1.19e-164		
Kurtosis:	12.308		Cond. No.	9.40e+10		

Model selanjutnya yang kami usulkan adalah penyederhanaan dari model 1. Penyederhanaan ini menggunakan metode variable screening dengan Teknik backward regression. Dengan Teknik ini, didapatkan model tanpa variable 'CHMIN'. Maka akan dihilangkan 'CHMIN' dan menggunakan variabel "MYCT", "MMIN", "MMAX", "CACH", dan "CHMAX" untuk membangun model regresi berikut ini.

$$Y = \beta_0 + \beta_1 MYCT + \beta_2 MMIN + \beta_3 MMAX + \beta_4 CACH + \beta_5 CHMAX + \epsilon; \epsilon \sim NIID(0, \sigma^2)$$

- β_0 adalah intercept (konstanta) dari model
- $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ adalah koefisien regresi untuk masing-masing variabel independen "MYCT", "MMIN", "MMAX", "CACH", dan "CHMAX"

$$\hat{y} = -56.075 + 0.0491MYCT + 0.0152MMIN + 0.0056MMAX + 0.6298CACH + 1.4599CHMIN$$

OLS Regression Results						
=====						
Dep. Variable:	PRP	R-squared:	0.865			
Model:	OLS	Adj. R-squared:	0.861			
Method:	Least Squares	F-statistic:	259.			
Date:	Sat, 16 Dec 2023	Prob (F-statistic):	3.86e-86			
Time:	14:44:22	Log-Likelihood:	-1148.7			
No. Observations:	209	AIC:	2309.			
Df Residuals:	203	BIC:	2330.			
Df Model:	5					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	-56.0750	8.007	-7.003	0.000	-71.862	-40.288
MYCT	0.0491	0.017	2.813	0.005	0.015	0.084
MMIN	0.0152	0.002	8.490	0.000	0.012	0.019
MMA	0.0056	0.001	8.695	0.000	0.004	0.007
CACH	0.6298	0.134	4.687	0.000	0.365	0.895
CHMAX	1.4599	0.208	7.031	0.000	1.050	1.869
=====						
Omnibus:	102.713	Durbin-Watson:	1.199			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1197.698			
Skew:	1.550	Prob(JB):	8.38e-261			
Kurtosis:	14.310	Cond. No.	3.31e+04			

3.7 Model 4 : Backward Selection dengan Interaksi dan Kuadrat

Model selanjutnya yang kami usulkan adalah model dengan screening metode backward, tetapi dengan model awal memasukkan semua interaksi antar predictor dan polinom orde 2 dari prediktor. Dengan Teknik ini, didapatkan model yang cukup kompleks (melibatkan 5 interaksi terms dan 1 orde 2).

$$Y = \beta_0 + \beta_1 MMIN + \beta_2 MMAX + \beta_3 CACH + \beta_4 CHMIN + \beta_5 CHMAX + \beta_6 CHMIN:CHMAX + \beta_7 MMAX:CACH + \beta_8 MMAX:CHMAX + \beta_9 MMIN:CACH + \beta_{10} MMIN:MMAX + \beta_{12} MMIN^2 + \epsilon ; \epsilon \sim NIID(0, \sigma^2)$$

- β_0 adalah intercept (konstanta) dari model
- $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ adalah koefisien regresi untuk masing-masing variabel independen "MMIN", "MMAX", "CACH", dan "CHMAX"
- β_6 koefisien untuk $CHMIN:CHMAX$, β_7 koefisien untuk $MMAX:CACH$, β_8 koefisien untuk $MMAX:CHMAX$, β_9 koefisien untuk $MMIN:CACH$, β_{10} koefisien untuk $MMIN:MMAX$, β_{12} koefisien untuk $MMIN^2$

$$\hat{y} = -25.3812 + 0.0097MMIN - 0.0043MMAX + 0.1806CACH - 1.4297CHMIN + 3.4017CHMAX + 0.0517MYCT - 0.0141CHMAX:MYCT + 7.152e - 05MMIN:CACH$$

OLS Regression Results

Dep. Variable:	PRP	R-squared:	0.907
Model:	OLS	Adj. R-squared:	0.903
Method:	Least Squares	F-statistic:	243.6
Date:	Mon, 18 Dec 2023	Prob (F-statistic):	1.00e-98
Time:	10:16:00	Log-Likelihood:	-1109.7
No. Observations:	209	AIC:	2237.
Df Residuals:	200	BIC:	2268.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	-25.3812	7.795	-3.256	0.001	-40.752	-10.011
MMIN	0.0097	0.002	4.452	0.000	0.005	0.014
MMAX	0.0043	0.001	7.476	0.000	0.003	0.005
CACH	0.1806	0.140	1.287	0.200	-0.096	0.457
CHMIN	-1.4297	0.730	-1.960	0.051	-2.868	0.009
CHMAX	3.4017	0.284	11.988	0.000	2.842	3.961
MYCT	0.0517	0.015	3.411	0.001	0.022	0.082
CHMAX:MYCT	-0.0141	0.002	-8.782	0.000	-0.017	-0.011
MMIN:CACH	7.152e-05	1.65e-05	4.329	0.000	3.89e-05	0.000

R^2 terlihat sangat baik, tetapi model juga sangat kompleks dengan total 13 parameter yang harus diestimasi (termasuk intercept).

BAB 4

Pengolahan Data dan Analisis Hasil

4.1 Uji Asumsi Model 1

4.1.1 Uji Normalitas

Shapiro-Wilk Test Statistic: 0.831267237663269

P-value: 2.5972681375941754e-14

The residuals do not appear to be normally distributed (reject H_0)

Karena asumsi normalitas dilanggar, maka akan diuji normalitas residual dengan menghilangkan outlier residualnya.

Shapiro-Wilk Test Statistic: 0.9603933691978455

P-value: 2.288530595251359e-05

The residuals do not appear to be normally distributed (reject H_0)

Asumsi normalitas masih dilanggar, maka model 1 akan ditransformasi dengan menggunakan \sqrt{y}

Shapiro-Wilk Test Statistic: 0.9912457466125488

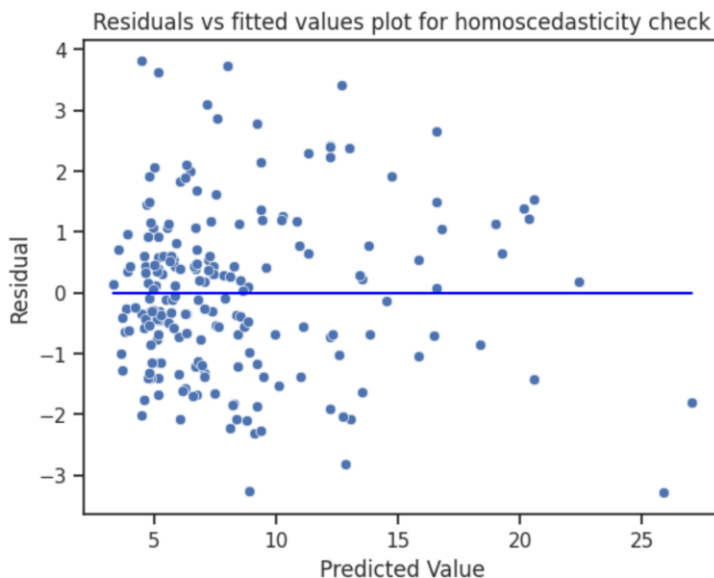
P-value: 0.253154456615448

The residuals appear to be normally distributed (fail to reject H_0)

Asumsi normalitas terpenuhi, maka model 1 yang akan dimasukkan sebagai kandidat adalah :

const	3.8834	$\hat{y}^* = \sqrt{y}$ $\hat{y}^* = 3.8834 - 0.0007MYCT + 0.0004MMIN + 0.0002MMAX + 0.0288CACH + 0.0225CHMIN + 0.0267CHMAX$
MYCT	-0.0007	
MMIN	0.0004	
MMAX	0.0002	
CACH	0.0288	
CHMIN	0.0225	
CHMAX	0.0267	

4.1.2 Homoscedasticity



Dapat dilihat, plot tidak menunjukkan pola tertentu.

4.1.3 Uji Goldfeld Quandt

Uji Goldfeld Quandt menguji heteroskedastisitas error model dengan hipotesis

H_0 : Suku error homoskedastic

H_1 : Suku error heteroskedastic

```
import statsmodels.stats.api as sms
from statsmodels.compat import lzip
name = ['F statistic', 'p-value']
test = sms.het_goldfeldquandt(model1.resid, X)
lzip(name, test)
```

```
[('F statistic', 0.8843169970718371), ('p-value', 0.7252534907444257)]
```

Karena $p\text{-value} > 0.05$, maka H_0 tidak dapat ditolak, asumsi homoskedastisitas terpenuhi.

4.1.4 Uji Independensi

```
from statsmodels.stats.stattools import durbin_watson as dwtest
def dwtestfunc(residual):
    print("H0 : residuals dari model independen")
    print("H1 : residuals dari model tidak independen")
    print("Titik Kritis 1,5< Dwtest < 2.5")
    if 1.5 < dwtest(resids=np.array(residual)) < 2.5:
        print("H0 diterima, artinya residual terbukti independen")
        print(dwtest(resids=np.array(residual)))
    else:
        print("H0 ditolak, artinya residual terbukti tidak independen")
        print(dwtest(resids=np.array(residual)))
dwtestfunc(residuals)
```

```
H0 : residuals dari model independen
H1 : residuals dari model tidak independen
Titik Kritis 1,5< Dwtest < 2.5
H0 ditolak, artinya residual terbukti tidak independen
1.4373087984539399
```

Setelah dilakukan Uji Durbin-Watson, didapatkan nilai statistik 1.44, dan asumsi independence of residuals tidak terpenuhi.

4.2 Uji Asumsi Model 2

4.2.1 Uji Normalitas

```
Shapiro-Wilk Test Statistic: 0.8512084484100342
P-value: 2.295042411323833e-13
The residuals do not appear to be normally distributed (reject H0)
```

Karena asumsi normalitas dilanggar, maka akan diuji normalitas residual dengan menghilangkan outlier residualnya.

```
Shapiro-Wilk Test Statistic: 0.9645893573760986
P-value: 6.416290125343949e-05
The residuals do not appear to be normally distributed (reject H0)
```

Asumsi normalitas masih dilanggar, maka model 2 akan ditransformasi dengan menggunakan $\log(y)$

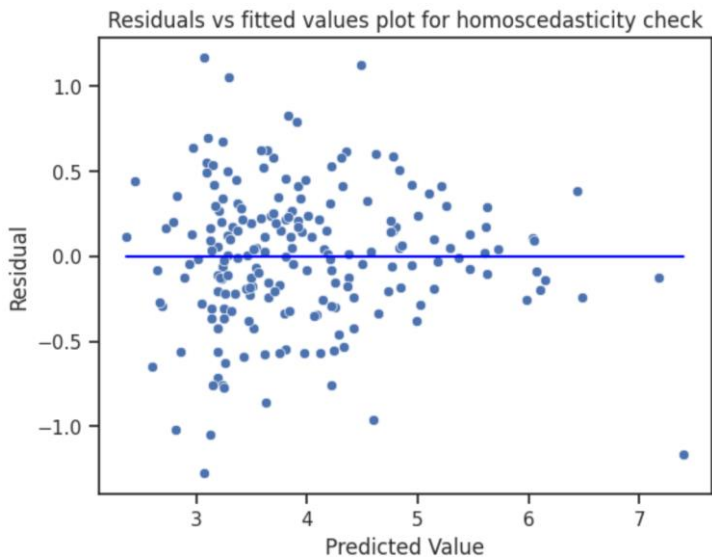
```
Shapiro-Wilk Test Statistic: 0.9921792149543762
P-value: 0.3428015112876892
```

The residuals appear to be normally distributed (fail to reject H_0)

Asumsi normalitas terpenuhi, maka model 2 yang akan dimasukkan sebagai kandidat adalah :

const	2.9937	$\hat{y}^* = \log(y)$
MYCT	-0.0005	$\hat{y}^* = 2.9937 - 0.0005MYCT + 0.00004714MMIN + 0.00007.738MMAX$
MMIN	4.714e-05	$+ 0.0072CACH - 0.0025CHMIN + 0.0056CHMAX$
MMAX	7.738e-05	$+ 4.8252\left(\frac{1}{MYCT}\right) - 0.0000000009009MMAX^2$
CACH	0.0072	
CHMIN	-0.0025	
CHMAX	0.0056	
1/MYCT	4.8252	
MMAX^2	-9.009e-10	

4.2.2 Homoskedasitas



Dapat dilihat, plot tidak menunjukkan pola tertentu.

4.2.3 Uji Goldfeld Quandt

Uji Goldfeld Quandt menguji heteroskedastisitas error model dengan hipotesis

H_0 : Suku error homoskedastic

H_1 : Suku error heteroskedastic

```
import statsmodels.stats.api as sms
from statsmodels.compat import lzip
name = ['F statistic', 'p-value']
test = sms.het_goldfeldquandt(model3.resid, X)
lzip(name, test)
```

```
[('F statistic', 1.0006078848569122), ('p-value', 0.4986836987957374)]
```

Karena p-value lebih besar dari 0.05, maka H_0 tidak dapat ditolak sehingga asumsi homoskedastisitas terpenuhi.

4.2.4 Uji Independensi

```
from statsmodels.stats.stattools import durbin_watson as dwtest
def dwtestfunc(residual):
    print("H0 : residuals dari model independen")
    print("H1 : residuals dari model tidak independen")
    print("Titik Kritis 1,5< Dwtest < 2.5")
    if 1.5 < dwtest(resids=np.array(residual)) < 2.5:
        print("H0 diterima, artinya residual terbukti independen")
        print(dwtest(resids=np.array(residual)))
    else:
        print("H0 ditolak, artinya residual terbukti tidak independen")
        print(dwtest(resids=np.array(residual)))
    dwtestfunc(residuals)
```

```
H0 : residuals dari model independen
H1 : residuals dari model tidak independen
Titik Kritis 1,5< Dwtest < 2.5
H0 ditolak, artinya residual terbukti tidak independen
1.4654382086826816
```

Setelah dilakukan Uji Durbin-Watson, didapatkan nilai statistik 1.46, dan asumsi independence of residuals tidak terpenuhi.

4.3 Uji Asumsi Model 3

4.3.1 Uji Normalitas

Shapiro-Wilk Test Statistic: 0.8300414085388184

P-value: 2.2853985685248462e-14

The residuals do not appear to be normally distributed (reject H_0)

Karena asumsi normalitas dilanggar, maka akan diuji normalitas residual dengan menghilangkan outlier residualnya.

Shapiro-Wilk Test Statistic: 0.9607220888137817

P-value: 2.4846322048688307e-05

The residuals do not appear to be normally distributed (reject H_0)

Asumsi normalitas masih dilanggar, maka model 3 akan ditransformasi dengan menggunakan \sqrt{y}

Shapiro-Wilk Test Statistic: 0.9937098622322083

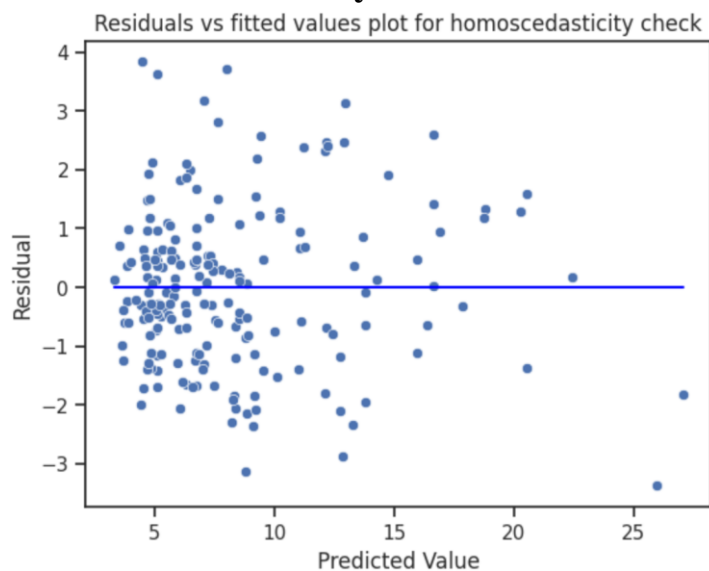
P-value: 0.5386074781417847

The residuals appear to be normally distributed (fail to reject H_0)

Asumsi normalitas terpenuhi, maka model 3 yang akan dimasukkan sebagai kandidat adalah :

const	3.8537	$\hat{y}^* = \sqrt{y}$ $\hat{y}^* = 3.8537 - 0.0007MYCT + 0.0005MMIN + 0.0002MMAX + 0.0297CACH + 0.0314CHMAX$
MYCT	-0.0007	
MMIN	0.0005	
MMAX	0.0002	
CACH	0.0297	
CHMAX	0.0314	

4.3.2 Homoscedasticity



Dapat dilihat, plot tidak menunjukkan pola tertentu.

4.3.3 Uji Goldfeld Quandt

Uji Goldfeld Quandt menguji heteroskedastisitas error model dengan hipotesis

H_0 : Suku error homoskedastik

H_1 : Suku error heteroskedastik

```

1 import statsmodels.stats.api as sms
2 from statsmodels.compat import lzip
3 name = ['F statistic', 'p-value']
4 test = sms.het_goldfeldquandt(residuals, X)
5 lzip(name, test)

```

```
[('F statistic', 0.9031795840454423), ('p-value', 0.6881107003265077)]
```

Karena p-value lebih besar dari 0.05, maka H_0 tidak dapat ditolak sehingga asumsi homoskedastisitas terpenuhi.

4.3.4 Uji Independensi

```

1 from statsmodels.stats.stattools import durbin_watson as dwtest
2 def dwtestfunc(residual):
3     print("H0 : residuals dari model independen")
4     print("H1 : residuals dari model tidak independen")
5     print("Titik Kritis 1,5< Dwtest < 2.5")
6     if 1.5 < dwtest(resids=np.array(residual)) < 2.5:
7         print("H0 ditolak, artinya residual terbukti tidak independen")
8         print(dwtest(resids=np.array(residual)))
9     else:
10        print("H0 tidak ditolak, artinya residual terbukti independen")
11        print(dwtest(resids=np.array(residual)))
12 dwtestfunc(residuals)

```

```

H0 : residuals dari model independen
H1 : residuals dari model tidak independen
Titik Kritis 1,5< Dwtest < 2.5
H0 tidak ditolak, artinya residual terbukti independen
1.458371844256196

```

Setelah dilakukan Uji Durbin-Watson, didapatkan nilai statistik 1.46, dan asumsi independence of residuals terpenuhi.

4.4 Uji Asumsi Model 4

4.4.1 Uji Normalitas

Shapiro-Wilk Test Statistic: 0.8885896801948547

P-value: 2.5202031433968486e-11

The residuals do not appear to be normally distributed (reject H_0)

Karena asumsi normalitas dilanggar, maka akan diuji normalitas residual dengan menghilangkan outlier residualnya.

Shapiro-Wilk Test Statistic: 0.9665036797523499

P-value: 8.634341793367639e-05

The residuals do not appear to be normally distributed (reject H_0)

Asumsi normalitas masih dilanggar, maka model 4 akan ditransformasi dengan menggunakan $\log(y)$

Shapiro-Wilk Test Statistic: 0.9921369552612305

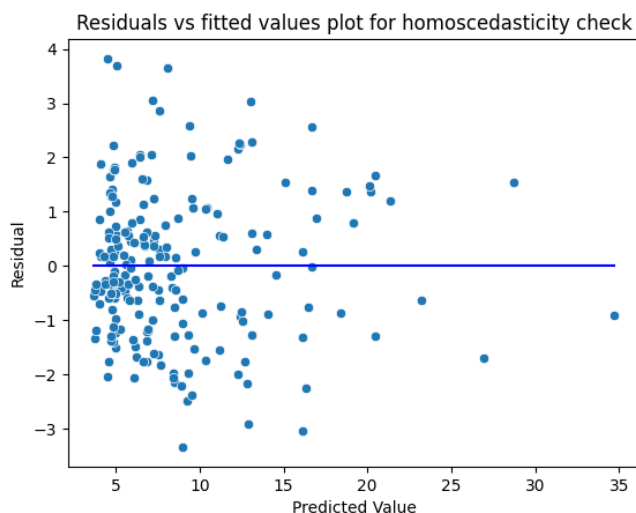
P-value: 0.33824819326400757

The residuals appear to be normally distributed (fail to reject H_0)

Asumsi normalitas terpenuhi, maka model 5 yang akan dimasukkan sebagai kandidat adalah :

const	3.8373	$\hat{y}^* = \log(y)$
MMIN	0.0005	$\hat{y}^* = 3.8373 + 0.0005MMIN + 0.0002MMAX + 0.0342CACH$
MMA	0.0002	$+ 0.0119CHMIN + 0.0350CHMAX - 0.0004MYCT$
CACH	0.0342	$- 0.0001CHMAX:MYCT$
CHMIN	0.0119	$- 0.000001147MMIN:CACH$
CHMAX	0.0350	
MYCT	-0.0004	
CHMAX:MYCT	-0.0001	
MMIN:CACH	-1.147e-06	

4.4.2 Homoscedasticity



Dapat dilihat, plot tidak menunjukkan pola tertentu.

4.4.3 Uji Goldfeld Quandt

Uji Goldfeld Quandt menguji heteroskedastisitas error model dengan hipotesis

H_0 : Suku error homoskedastik

H_1 : Suku error heteroskedastik

```
import statsmodels.stats.api as sms
from statsmodels.compat import lzip
name = ['F statistic', 'p-value']
test = sms.het_goldfeldquandt(model5.resid, X)
lzip(name, test)
```

```
[('F statistic', 0.8532263542320913), ('p-value', 0.7777116265985632)]
```


Karena p-value lebih besar dari 0.05, maka H_0 tidak dapat ditolak sehingga asumsi homoskedastisitas terpenuhi.

4.4.4 Uji Independensi

```
from statsmodels.stats.stattools import durbin_watson as dwtest
def dwtestfunc(residual):
    print("H0 : residuals dari model independen")
    print("H1 : residuals dari model tidak independen")
    print("Titik Kritis 1,5< Dwtest < 2.5")
    if 1.5 < dwtest(resids=np.array(residual)) < 2.5:
        print("H0 ditolak, artinya residual terbukti tidak independen")
        print(dwtest(resids=np.array(residual)))
    else:
        print("H0 tidak ditolak, artinya residual terbukti independen")
        print(dwtest(resids=np.array(residual)))
dwtestfunc(residual)
```

```
H0 : residuals dari model independen
H1 : residuals dari model tidak independen
Titik Kritis 1,5< Dwtest < 2.5
H0 tidak ditolak, artinya residual terbukti independen
1.3638015879052532
```

Setelah dilakukan Uji Durbin-Watson, didapatkan nilai statistik 1.3638, dan asumsi independence of residuals tidak terpenuhi.

BAB 5

Penutup

Berikut beberapa informasi yang didapatkan melalui analisis, yaitu:

- Dengan metode forward selection, kita dapatkan $Y = \beta_0 + \beta_1 MYCT + \beta_2 MMIN + \beta_3 MMAX + \beta_4 CACH + \beta_5 CHIN + \beta_6 CHMAX + \epsilon ; \epsilon \sim NIID(0, \sigma^2)$ (model 1) dan dengan backward selection kita dapatkan (model 3) $Y = \beta_0 + \beta_1 CACH + \beta_2 CHMAX + \beta_3 MMAX + \beta_4 MMIN + \beta_5 MYCT$
- Dengan melihat pairplot, kita tentukan bahwa ada hubungan terbalik antara MYCT dan PRP dan ada curvature antara interaksi PRP dan MMAX. Dari kedua observasi tersebut, kita tentukan (model 2) $Y = \beta_0 + \beta_1 MYCT + \beta_2 MMIN + \beta_3 MMAX + \beta_4 CACH + \beta_5 CHIN + \beta_6 CHMAX + \beta_7 \left(\frac{1}{MYCT}\right) + \beta_8 MMAX^2 + \epsilon ; \epsilon \sim NIID(0, \sigma^2)$ sebagai salah satu model yang kita usulkan.
- Kita juga melakukan metode backward selection, dengan complete modelnya adalah model orde 2 dengan interaksi dan mendapatkan (model 4) $Y = \beta_0 + \beta_1 MMIN + \beta_2 MMAX + \beta_3 CACH + \beta_4 CHMIN + \beta_5 CHMAX + \beta_6 CHMIN:CHMAX + \beta_7 MMAX:CACH + \beta_8 MMAX:CHMAX + \beta_9 MMIN:CACH + \beta_{10} MMIN:MMAX + \beta_{12} MMIN^2 + \epsilon ; \epsilon \sim NIID(0, \sigma^2)$

Lalu, Kita Uji asumsi dan mendapatkan beberapa model melanggar asumsi normalitas dan independence atau autocorrelation. Kita menangani masalah normalitas dengan mentransformasi target atau variable PRP pada model tersebut dan mendapatkan hasil akhir pada setiap sebagai berikut:

Model 1

$$\widehat{y^*} = \sqrt{y}$$

$$\widehat{y^*} = \beta_0 + \beta_1 MYCT + \beta_2 MMIN + \beta_3 MMAX + \beta_4 CACH + \beta_5 CHIN + \beta_6 CHMAX + \epsilon ; \epsilon \sim NIID(0, \sigma^2)$$

Model 2

$$\widehat{y^*} = \log(y)$$

$$\widehat{y^*} = \beta_0 + \beta_1 MYCT + \beta_2 MMIN + \beta_3 MMAX + \beta_4 CACH + \beta_5 CHIN + \beta_6 CHMAX + \beta_7 \left(\frac{1}{MYCT}\right) + \beta_8 MMAX^2 + \epsilon ; \epsilon \sim NIID(0, \sigma^2)$$

Model 3

$$\widehat{y^*} = \sqrt{y}$$

$$\widehat{y^*} = \beta_0 + \beta_1 MYCT + \beta_2 MMIN + \beta_3 MMAX + \beta_4 CACH + \beta_5 CHMAX + \epsilon ; \epsilon \sim NIID(0, \sigma^2)$$

Model 4

$$\widehat{y^*} = \sqrt{y}$$

$$\widehat{y^*} = \beta_0 + \beta_1 MMIN + \beta_2 MMAX + \beta_3 CACH + \beta_4 CHMIN + \beta_5 CHMAX + \beta_6 CHMIN:CHMAX + \beta_7 MMAX:CACH + \beta_8 MMAX:CHMAX + \beta_9 MMIN:CACH + \beta_{10} MMIN:MMAX + \beta_{12} MMIN^2 + \epsilon ; \epsilon \sim NIID(0, \sigma^2)$$

Dengan Hasil Regresi setelah transformasi sebagai berikut : (dari Python)

Model 1

OLS Regression Results						
=====						
Dep. Variable:	sqrtPRP	R-squared:	0.904			
Model:	OLS	Adj. R-squared:	0.901			
Method:	Least Squares	F-statistic:	310.9			
Date:	Tue, 19 Dec 2023	Prob (F-statistic):	6.87e-98			
Time:	03:00:42	Log-Likelihood:	-372.70			
No. Observations:	205	AIC:	759.4			
Df Residuals:	198	BIC:	782.7			
Df Model:	6					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	3.9160	0.216	18.170	0.000	3.491	4.341
MYCT	-0.0008	0.000	-1.788	0.075	-0.002	8.27e-05
MMIN	0.0004	5.27e-05	8.001	0.000	0.000	0.001
MMAX	0.0002	1.69e-05	12.187	0.000	0.000	0.000
CACH	0.0280	0.004	7.590	0.000	0.021	0.035
CHMIN	0.0375	0.023	1.655	0.100	-0.007	0.082
CHMAX	0.0230	0.006	3.596	0.000	0.010	0.036
=====						
Omnibus:	22.696	Durbin-Watson:	1.569			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	46.723			
Skew:	0.529	Prob(JB):	7.15e-11			
Kurtosis:	5.086	Cond. No.	3.24e+04			
=====						

Model 2

OLS Regression Results						
Dep. Variable:	logPRP	R-squared:	0.839			
Model:	OLS	Adj. R-squared:	0.832			
Method:	Least Squares	F-statistic:	127.6			
Date:	Tue, 19 Dec 2023	Prob (F-statistic):	1.92e-73			
Time:	03:00:44	Log-Likelihood:	-105.73			
No. Observations:	205	AIC:	229.5			
Df Residuals:	196	BIC:	259.4			
Df Model:	8					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	3.0084	0.087	34.595	0.000	2.837	3.180
MYCT	-0.0005	0.000	-3.666	0.000	-0.001	-0.000
MMIN	4.72e-05	1.44e-05	3.267	0.001	1.87e-05	7.57e-05
MMAX	7.98e-05	8e-06	9.978	0.000	6.4e-05	9.56e-05
CACH	0.0065	0.001	6.142	0.000	0.004	0.009
CHMIN	-0.0009	0.006	-0.151	0.880	-0.013	0.011
CHMAX	0.0046	0.002	2.567	0.011	0.001	0.008
1/MYCT	3.9970	4.149	0.963	0.337	-4.185	12.179
MMAX^2	-8.98e-10	1.68e-10	-5.346	0.000	-1.23e-09	-5.67e-10
Omnibus:	3.426	Durbin-Watson:	1.465			
Prob(Omnibus):	0.180	Jarque-Bera (JB):	3.634			
Skew:	-0.120	Prob(JB):	0.163			
Kurtosis:	3.606	Cond. No.	7.87e+10			

Model 3

OLS Regression Results						
Dep. Variable:	squaredPRP	R-squared:	0.903			
Model:	OLS	Adj. R-squared:	0.900			
Method:	Least Squares	F-statistic:	369.4			
Date:	Tue, 19 Dec 2023	Prob (F-statistic):	1.32e-98			
Time:	03:00:46	Log-Likelihood:	-374.11			
No. Observations:	205	AIC:	760.2			
Df Residuals:	199	BIC:	780.2			
Df Model:	5					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	3.9223	0.216	18.123	0.000	3.496	4.349
MYCT	-0.0008	0.000	-1.821	0.070	-0.002	6.81e-05
MMIN	0.0004	5.26e-05	8.219	0.000	0.000	0.001
MMAX	0.0002	1.69e-05	12.382	0.000	0.000	0.000
CACH	0.0294	0.004	8.158	0.000	0.022	0.036
CHMAX	0.0272	0.006	4.629	0.000	0.016	0.039
Omnibus:	21.895	Durbin-Watson:	1.562			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	50.036			
Skew:	0.469	Prob(JB):	1.36e-11			
Kurtosis:	5.231	Cond. No.	3.24e+04			

Model 4

OLS Regression Results						
=====						
Dep. Variable:	modPRP		R-squared:	0.930		
Model:	OLS		Adj. R-squared:	0.927		
Method:	Least Squares		F-statistic:	326.7		
Date:	Tue, 19 Dec 2023		Prob (F-statistic):	6.34e-109		
Time:	03:00:48		Log-Likelihood:	-351.68		
No. Observations:	205		AIC:	721.4		
Df Residuals:	196		BIC:	751.3		
Df Model:	8					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	3.8373	0.215	17.859	0.000	3.414	4.261
MMIN	0.0005	6.05e-05	8.209	0.000	0.000	0.001
MMAX	0.0002	1.6e-05	12.605	0.000	0.000	0.000
CACH	0.0342	0.004	8.034	0.000	0.026	0.043
CHMIN	0.0119	0.021	0.557	0.578	-0.030	0.054
CHMAX	0.0350	0.009	3.834	0.000	0.017	0.053
MYCT	-0.0004	0.000	-0.990	0.323	-0.001	0.000
CHMAX:MYCT	-0.0001	4.82e-05	-2.771	0.006	-0.000	-3.85e-05
MMIN:CACH	-1.147e-06	4.64e-07	-2.471	0.014	-2.06e-06	-2.32e-07
=====						
Omnibus:	2.780		Durbin-Watson:	1.364		
Prob(Omnibus):	0.249		Jarque-Bera (JB):	2.734		
Skew:	0.281		Prob(JB):	0.255		
Kurtosis:	2.930		Cond. No.	1.09e+06		
=====						

Perbandingan Model

	Model	rsquared	rsquared_adj	AIC	BIC	MSE	Press	Parameter
0	model1	0.904055	0.901148	759.404267	782.665337	2.300173	522.147539	6
1	model2	0.838929	0.832355	229.465194	259.372284	0.171797	39.285599	8
2	model3	0.902729	0.900285	760.219038	780.157098	2.320255	518.420634	5
3	model4	0.930240	0.927393	721.366944	751.274034	1.892837	432.184003	8

Dapat dilihat dari tabel perbandingan diatas didapatkan bahwa model 2 adalah model terbaik dengan alasan memiliki AIC,BIC,MSE, dan Press terbaik diantara 4 model.

Model 2 :

$$\widehat{y^*} = \log(y)$$

$$\widehat{y^*} = \beta_0 + \beta_1 MYCT + \beta_2 MMIN + \beta_3 MMAX + \beta_4 CACH + \beta_5 CHIN + \beta_6 CHMAX + \beta_7 \left(\frac{1}{MYCT} \right) + \beta_8 MMAX^2 + \epsilon ; \epsilon \sim NIID(0, \sigma^2)$$

Tetapi sulit untuk mengatakan model 2 terbaik secara absolut dikarenakan perbedaan transformasi pada model1,model3 dan model4(\sqrt{y}) dengan model2($\log(y)$). Jadi, kita dapat memilih juga model terbaik pada transformasi \sqrt{y} .

Jika memilih best model dengan transformasi \sqrt{y} , kita dapatkan model4 sebagai model terbaik dengan rsquare,rsquared_adj,AIC,BIC,MSE,Press terkecil di antara ketiga model dengan transformasi \sqrt{y} .

Model 4:

$$\widehat{y^*} = \sqrt{y}$$

$$\widehat{y^*} = \beta_0 + \beta_1 MMIN + \beta_2 MMAX + \beta_3 CACH + \beta_4 CHMIN + \beta_5 CHMAX + \beta_6 CHMIN: CHMAX + \beta_7 MMAX: CACH + \beta_8 MMAX: CHMAX + \beta_9 MMIN: CACH + \beta_{10} MMIN: MMAX + \beta_{12} MMIN^2 + \epsilon ; \epsilon \sim NIID(0, \sigma^2)$$

Jika menitik beratkan pada prinsip Parsimonious, kita bisa memiliki model3 dikarenakan memiliki Rsquared,Rsquared_adj,AIC,BIC,MSE dan Press yang tidak terlalu berbeda dengan Model4 tetapi memiliki parameter yang lebih dikit.

Note

Pada keempat model, kita temukan variable dengan p-value yang lebih besar dari $\alpha = 0.05$ ini dikarenakan sifat dari forward selection yang memiliki prinsip jika sudah ada variable yang diterima tidak akan dibuang. P-value yang lebih dari $\alpha = 0.05$ juga disebabkan oleh transformasi variable response dengan tujuan menormalkan residual.

Pada keempat model, kita juga temukan condition number yang tinggi yang menunjukkan bahwa ada masalah kestabilan diantara koefisien variable kita yang dapat di sebabkan oleh masa satuan yang memberi memberikan efek yang tidak stabil antar variable. (Condition number yang besar juga bisa disebabkan oleh multikolinearitas tetapi sudah kita buktikan mematuhi asumsi)

BAB 6

Lampiran

Link codes Python : [Molin_Project2_Regression.ipynb - Colaboratory \(google.com\)](#)