Analisis Faktor-Faktor yang Mempengaruhi Durasi Tidur Menggunakan Model Regresi Linear

Makalah ini disusun untuk memenuhi Project Mata Kuliah Model Linear

Penerapan Metode Regresi Linear pada Dataset yang Relevan

Dosen Pengampu : Madona Yunita Wijaya, M.Sc.

Disusun oleh:

Ahmad Farhan Sarofi	11230940000006
Hardina Dewita	11230940000014
Ghazalah Carissa Qathilah	11230940000026
Galih Pungkas Puruwito	11230940000058

PROGRAM STUDI MATEMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS ISLAM NEGERI SYARIF HIDAYATULLAH JAKARTA

2025

ABSTRAK

KATA PENGANTAR

Assalamualaikum Warahmatullahi Wabarakatuh

Alhamdulillah, puji syukur kita panjatkan kepada Allah SWT. karena berkat rahmat dan hidayah-Nya kita dapat menyelesaikan Laporan Projek Akhir untuk Mata Kuliah Model Linear yang berjudul **"Analisis Faktor-Faktor yang Mempengaruhi Durasi Tidur Menggunakan Model Regresi Linear"**. Shalawat serta salam tidak lupa kita panjatkan kepada junjungan nabi besar Nabi Muhammad SAW. beserta keluarganya, para sahabat, dan para pengikutnya.

Laporan Projek Akhir ini merupakan salah satu kewajiban yang harus dilaksanakan oleh mahasiswa yang mengikuti Mata Kuliah Model Linear Program Studi Matematika Fakultas Sains dan Teknologi UIN Syarif Hidayatullah Jakarta sebagai pengimplementasian teori yang sudah dipelajari selama perkuliahan berlangsung.

Kami menyadari bahwa dalam penulisan makalah ini tidak terlepas dari bantuan banyak pihak. Oleh karena itu, kami mengucapkan terima kasih dan mohon maaf apabila ada kekurangan dalam penulisan laporan ini.

Wassalamualaikum Warahmatullahi Wabarakatuh

Jakarta, 10 Juli 2025

Penulis

DAFTAR ISI

BAB I

PENDAHULUAN

1.1. Latar Belakang

Di tengah gaya hidup modern yang serba cepat dan penuh tekanan, kualitas serta durasi tidur menjadi aspek krusial dalam menjaga keseimbangan kesehatan fisik maupun mental. Kurangnya waktu tidur yang cukup telah terbukti berhubungan dengan berbagai masalah kesehatan, seperti gangguan metabolik, penyakit kardiovaskular, gangguan kognitif, hingga penurunan performa dalam aktivitas sehari-hari. Oleh sebab itu, memahami faktor-faktor yang memengaruhi durasi tidur menjadi penting, terutama dalam konteks pencegahan dan intervensi berbasis data empiris.

Dalam ranah analisis kuantitatif, regresi linear merupakan metode yang umum digunakan untuk menganalisis hubungan antara satu variabel respon dengan satu atau lebih variabel prediktor. Pada penelitian ini, pendekatan regresi linear dimanfaatkan untuk mengevaluasi pengaruh sejumlah faktor gaya hidup dan karakteristik individu, termasuk usia, kualitas tidur, tingkat stres, denyut jantung, tingkat aktivitas fisik, jenis kelamin, dan kategori BMI, terhadap durasi tidur. Melalui teknik ini, dimungkinkan untuk membangun model prediktif yang mampu mengidentifikasi besarnya kontribusi masing-masing faktor terhadap variasi durasi tidur.

Akan tetapi, validitas hasil dari model regresi sangat ditentukan oleh sejauh mana asumsi dasar regresi dipenuhi, seperti linearitas, normalitas residual, homoskedastisitas, dan independensi error. Jika asumsi-asumsi ini dilanggar, maka keakuratan dan interpretasi model dapat terpengaruh. Oleh karena itu, penelitian ini turut melakukan pemeriksaan diagnostik model, termasuk transformasi variabel, pengujian multikolinearitas, dan penerapan metode alternatif seperti Weighted Least Squares (WLS) serta validasi silang k-fold untuk meningkatkan keandalan dan generalisasi model yang dihasilkan.

Oleh karena itu, penelitian ini bertujuan untuk mengidentifikasi faktor-faktor signifikan yang memengaruhi durasi tidur dan mengembangkan model regresi yang representatif serta dapat dijadikan dasar dalam penyusunan kebijakan kesehatan atau penelitian lebih lanjut di bidang statistik terapan.

1.2. Rumusan Masalah

1. Faktor-faktor apa saja yang secara signifikan mempengaruhi variabel respon yang diteliti dan berapa besarnya?
2. Bagaimana performa model untuk prediksi?

1.3. Tujuan

Penelitian ini memiliki tujuan untuk:

1. Menelusuri dan mengkaji variabel-variabel independen yang berpengaruh secara signifikan terhadap variabel dependen dalam kerangka model statistik yang dibangun.
2. Mengestimasi besar pengaruh masing-masing variabel independen terhadap perubahan pada variabel dependen, baik secara individual maupun bersamaan.
3. Merancang model statistik yang secara representatif menggambarkan keterkaitan antara variabel dependen dan faktor-faktor penjelasnya.
4. Mengkaji performa model yang dihasilkan melalui analisis ketepatan prediksi serta kelayakan model berdasarkan indikator statistik yang sesuai.

1.4. Manfaat

Dari sisi teoretis, hasil penelitian ini diharapkan dapat memberikan kontribusi sebagai berikut:

1. Menambah wawasan teoretis mengenai pola hubungan antar variabel dalam sistem tertentu melalui pendekatan analisis kuantitatif.
2. Mendukung pengembangan ilmu statistik terapan, terutama dalam hal pemodelan regresi dan teknik prediksi berbasis data empiris.
3. Memberikan referensi akademik bagi studi-studi lanjutan yang berkaitan dengan pemodelan hubungan antar variabel atau evaluasi model statistik.

BAB II

KAJIAN TEORI

2.1 Eksplorasi Data

Eksplorasi data merupakan tahap awal yang sangat penting dalam proses analisis statistik karena memberikan pemahaman menyeluruh terhadap struktur dan distribusi data. Proses ini bertujuan untuk mengidentifikasi pola, anomali, atau potensi masalah yang dapat memengaruhi kualitas hasil analisis regresi.

1. Statistik Deskriptif

   Statistik deskriptif adalah langkah untuk mendapatkan gambaran umum tentang data seperti nilai rata-rata, median, dan nilai minimum/maksimum. Ini membantu kita memahami karakteristik dasar dari data yang kita miliki. Beberapa Metode Statistik Deskriptif:

   - Mean (Rata-rata): Ukuran pemusatan data yang diperoleh dengan menjumlahkan seluruh nilai dalam suatu kumpulan data, kemudian dibagi dengan jumlah banyaknya data.
   - Median: Nilai yang berada tepat di tengah kumpulan data setelah data tersebut diurutkan dari yang terkecil hingga terbesar.
   - Modus: Nilai yang paling sering muncul dalam data.
   - Maximum (Maksimum): Menunjukkan nilai terbesar dalam distribusi suatu variabel.
   - Minimum (Minimal): Menunjukkan nilai terkecil dalam distribusi suatu variabel.
   - Standard Deviation (Standar Deviasi): Akar kuadrat dari variansi, menunjukkan seberapa tersebar data di sekitar rata-rata.
   - Variance (Variansi): Ukuran dari seberapa jauh data tersebar dari rata-rata.

   Di R, fungsi summary() dapat memberikan statistik deskriptif untuk setiap kolom dalam dataframe. Fungsi ini membantu untuk mendapatkan nilai-nilai seperti mean, median, minimum, maksimum, dan quartiles.

2. Visualisasi Data

   Visualisasi data adalah langkah penting untuk memahami pola dan hubungan dalam data. Beberapa teknik visualisasi umum meliputi:

- Boxplot: Memberikan gambaran ringkas dari data melalui lima ukuran statistik penting, yaitu nilai minimum, kuartil pertama (Q1), median, kuartil ketiga (Q3), dan nilai maksimum. Grafik ini juga efektif dalam mendeteksi keberadaan nilai pencilan (outlier).
- Histogram: Digunakan untuk menggambarkan penyebaran frekuensi dari suatu variabel tunggal. Grafik ini membantu dalam mengenali bentuk distribusi data, seperti apakah data simetris, miring (skewed), atau memiliki outlier.
- Scatter Plot: Menunjukkan hubungan antara dua variabel numerik. Dengan grafik ini, kita dapat mengamati pola, kecenderungan, serta apakah terdapat hubungan linear atau non-linear antar variabel.
- Pair Plot (Scatterplot Matrix): Menampilkan grafik pencar untuk setiap pasangan kombinasi variabel dalam satu dataset. Di R, visualisasi ini dapat dibuat menggunakan fungsi pairs(). Plot ini sangat berguna untuk mengevaluasi korelasi antar variabel dalam skala yang lebih luas.

Visualisasi ini membantu menilai apakah asumsi linearitas dan normalitas dapat terpenuhi sebelum model regresi dibangun.

2.2 Uji Multikolinearitas

Multikolinearitas adalah kondisi di mana dua atau lebih variabel prediktor dalam model regresi memiliki korelasi tinggi, yang dapat menyebabkan ketidakstabilan dalam estimasi koefisien regresi. Masalah ini diidentifikasi menggunakan *Variance Inflation Factor* (VIF). VIF mengukur seberapa besar varians dari estimasi koefisien meningkat akibat adanya multikolinearitas.

Dalam penelitian ini, perhitungan VIF dilakukan pada model awal yang mengandung banyak interaksi dan transformasi. Beberapa variabel menunjukkan nilai VIF tinggi, mengindikasikan korelasi yang signifikan antar prediktor. Untuk mengatasi hal ini, dilakukan penyederhanaan model melalui seleksi variabel bertahap hingga diperoleh model akhir dengan VIF seluruh variabel < 10. Hal ini memastikan bahwa setiap prediktor memberikan kontribusi informasi unik terhadap model.

Beberapa cara untuk mengatasi multikolinearitas:

1. Menghapus salah satu variabel: Jika dua variabel sangat berkorelasi, salah satu dapat dihapus agar tidak menyebabkan redudansi informasi.

2. Menggabungkan variabel: Melalui teknik seperti *Principal Component Analysis* (PCA), variabel yang berkorelasi tinggi dapat digabungkan menjadi satu komponen baru yang independen.
3. Transformasi data: Transformasi log atau z-score terkadang dapat mengurangi korelasi antar variabel.
4. Menggunakan regresi *ridge* atau *penalized regression*: Model seperti *ridge regression* atau LASSO digunakan dalam regresi reguler untuk mengatasi multikolinearitas dengan menambahkan penalti terhadap besar koefisien.
5. Mengurangi jumlah variabel interaksi dan kuadrat yang tidak signifikan: Seperti yang dilakukan dalam penelitian ini, yaitu menghilangkan variabel interaksi dan polinomial yang tidak signifikan dapat menurunkan VIF dan menyederhanakan model.

Dengan menerapkan salah satu atau kombinasi dari strategi tersebut, stabilitas model dapat ditingkatkan dan interpretasi koefisien regresi menjadi lebih *reliable*.

2.3 Seleksi Prediktor

Seleksi prediktor adalah proses untuk memilih variabel input (independen) yang paling signifikan atau berkontribusi besar dalam menjelaskan variasi variabel dependen. Tujuan utama dari seleksi ini adalah meningkatkan kinerja model dengan mengurangi kompleksitas, menghindari *overfitting*, serta menghilangkan variabel yang tidak relevan atau bersifat redundan. Dalam penelitian ini, dua pendekatan utama yang digunakan adalah seleksi *stepwise* dan evaluasi subset terbaik.

1. Seleksi Prediktor Stepwise

   Seleksi *stepwise* merupakan metode otomatis yang menggabungkan pendekatan *forward selection* dan *backward elimination* menggunakan kriteria pemilihan *Akaike Information Criterion* (AIC). Prosedur ini diimplementasikan menggunakan fungsi *step()* dalam R. Tiga skenario algoritma yang umum digunakan:

   a. Algoritma *Backward elimination*: Dimulai dari model lengkap (*full model*) yang mencakup seluruh variabel, lalu menghapus satu per satu prediktor yang kontribusinya tidak signifikan jika penghapusannya dapat menurunkan nilai AIC.

b. Algoritma *Forward selection*: Dimulai dari model kosong, lalu menambahkan prediktor satu per satu yang memberikan penurunan AIC terbesar.
c. Algoritma *Both direction* (dua arah): Kombinasi *backward* dan *forward selection*. Pada setiap langkah, model dapat menambah atau menghapus prediktor untuk mencari nilai AIC terendah secara iteratif.

Dalam penelitian ini, seleksi *stepwise* dilakukan dengan *direction* = "both", dimulai dari model penuh yang mencakup semua variabel utama, kuadrat, dan interaksi. Hasil akhirnya menunjukkan bahwa hanya sebagian interaksi dan transformasi yang berkontribusi signifikan terhadap prediksi durasi tidur, sehingga disisakan dalam model akhir.

2. Seleksi *Best Subset*

Selain *stepwise*, pendekatan evaluasi subset terbaik digunakan untuk mempertimbangkan semua kombinasi variabel secara eksploratif. Meskipun pendekatan ini tidak dijalankan secara eksplisit dalam *script* R, prinsipnya digunakan dalam proses evaluasi model berdasarkan beberapa kriteria seleksi model terbaik, antara lain:

a. *Adjusted R-squared* (Adj-R^2): Mengukur proporsi variasi yang dijelaskan oleh model, dengan memperhitungkan jumlah prediktor. Cocok untuk membandingkan model dengan jumlah prediktor yang berbeda.
b. Mallows' Cp: Menilai seberapa baik model mendekati model "sebenarnya". Nilai Cp yang mendekati jumlah prediktor menunjukkan model yang efisien.
c. AIC dan BIC: Kriteria informasi yang mempertimbangkan *goodness-of-fit* dan kompleksitas model. Nilai AIC/BIC yang lebih kecil mengindikasikan model yang lebih baik.
d. PRESS (*Predicted Residual Sum of Squares*): Mengukur kemampuan prediksi model terhadap data baru. Semakin kecil PRESS, semakin baik performa prediksi model.

Dengan kombinasi pendekatan ini, model akhir diperoleh berdasarkan keseimbangan antara kompleksitas model, signifikansi statistik, dan kemampuan generalisasi melalui validasi silang 5-fold menggunakan package caret di R.

2.4 Model Regresi Linear

Regresi linear merupakan teknik statistik yang digunakan untuk memodelkan dan menganalisis hubungan antara satu variabel dependen (Y) dengan satu atau lebih variabel independen (X). Tujuannya adalah untuk memprediksi nilai variabel dependen berdasarkan nilai variabel independen serta untuk memahami seberapa besar kontribusi masing-masing prediktor terhadap variabel respon. Dalam konteks penelitian ini, regresi linear digunakan untuk memodelkan durasi tidur (*Sleep Duration*) berdasarkan faktor-faktor seperti usia (*age*), kualitas tidur (*quality of sleep*), aktivitas fisik (*physical activity level*), tingkat stres (*stress level*), denyut jantung (*heart rate*), jenis kelamin (*gender*), dan kategori BMI (*BMI category*).

1. Asumsi Model Regresi Linear
    a. Normalitas Residual: Mengasumsikan bahwa residual (kesalahan prediksi) berdistribusi normal. Asumsi ini penting karena banyak uji inferensial, termasuk uji signifikansi koefisien dan pembentukan interval kepercayaan, bergantung pada distribusi normal dari residual.
    b. Mean Residual = 0: Residual harus memiliki rata-rata mendekati nol. Ini menunjukkan bahwa prediksi model tidak bias secara sistematis.
    c. Homoskedastisitas: Varians dari residual harus konstan di seluruh rentang nilai prediktor. Pelanggaran terhadap asumsi ini disebut heteroskedastisitas, yang dapat menyebabkan kesalahan standar yang tidak akurat dan kesimpulan statistik yang menyesatkan.
    d. Independensi Residual: Residual harus independen satu sama lain, artinya tidak terdapat pola atau autokorelasi dalam sisa prediksi. Ini penting terutama ketika data diurutkan berdasarkan waktu atau kelompok.
2. Evaluasi Asumsi Model Regresi Linear
    a. Normalitas Residual
        - *Histogram dan Q-Q Plot*: Digunakan untuk melihat kesesuaian distribusi residual terhadap distribusi normal. Dalam penelitian ini, histogram

menunjukkan bentuk mendekati normal meski terdapat sedikit *skew* ke kiri, sementara Q-Q plot menunjukkan deviasi ringan pada ekor distribusi.

- *Shapiro-Wilk dan Kolmogorov-Smirnov Test*: Digunakan untuk menguji secara formal apakah residual berdistribusi normal. Hasil uji menunjukkan *p-value* mendekati atau lebih besar dari 0.05, yang mengindikasikan residual tidak berbeda signifikan dari distribusi normal.

b. Mean Residual = 0: Rata-rata residual dihitung secara eksplisit dan ditemukan sangat dekat dengan nol, sehingga asumsi ini terpenuhi.

c. Homoskedastisitas

- *Plot Residual vs Fitted Values*: Digunakan untuk melihat apakah terdapat pola tertentu dalam penyebaran residual. Hasil menunjukkan penyebaran yang relatif acak, meskipun terdapat sedikit pola menyebar.
- *Breusch-Pagan Test*: Uji formal terhadap homoskedastisitas menunjukkan adanya sedikit indikasi heteroskedastisitas ($p < 0.05$), namun tidak cukup kuat untuk mendiskualifikasi model.

d. Independensi Residual

- Durbin-Watson Test: Digunakan untuk mendeteksi autokorelasi residual. Nilai uji yang mendekati 2 menunjukkan bahwa tidak terdapat autokorelasi signifikan dalam data.

Model akhir yang digunakan dalam penelitian ini mencakup variabel transformasi kuadrat I(*Age*^2), I(*Stress Level*^2), I(*Heart Rate*^2), interaksi *Age* ; *Stress Level*, *Physical Activity Level* ; *Stress Level*, serta dua prediktor kategorik *Gender* dan BMI *Category*. Pemilihan variabel ini didasarkan pada hasil seleksi I dan validasi silang, dengan mempertimbangkan signifikansi statistik, multikolinearitas, serta evaluasi asumsi regresi.

2.5 Analisis Residual

Analisis residual adalah teknik statistik yang digunakan untuk mengevaluasi validitas dan kualitas model regresi. Residual merupakan selisih antara nilai aktual (observasi) dengan nilai yang diprediksi oleh model. Tujuan utama dari analisis residual

adalah untuk mendeteksi apakah terdapat pola atau struktur tertentu yang menandakan pelanggaran terhadap asumsi klasik regresi linear.

1. Metode Grafis dalam Analisis Residual
    a. *Quantile-Quantile Plot* (QQ Plot): Digunakan untuk memeriksa asumsi normalitas residual. Jika residual terdistribusi normal, maka titik-titik pada Q-Q plot akan mengikuti garis diagonal. Dalam penelitian ini, terdapat sedikit penyimpangan pada ekor distribusi, yang menandakan deviasi ringan dari normalitas.
    b. *Scatterplot Residual* vs. *Fitted Values*: Digunakan untuk mengecek homoskedastisitas dan mendeteksi *outlier*. Plot residual yang menyebar acak tanpa pola tertentu mengindikasikan homoskedastisitas. Dalam data ini, terlihat sedikit pola menyebar, yang menunjukkan kemungkinan heteroskedastisitas ringan.
    c. Histogram dan Boxplot: Histogram digunakan untuk melihat bentuk distribusi residual, sedangkan boxplot membantu mengidentifikasi pencilan (*outlier*) dan asimetri distribusi. Histogram menunjukkan bahwa residual cenderung mendekati normal, meskipun sedikit condong ke kiri.
2. Uji Signifikansi dalam Analisis Residual
    a. Uji Normalitas
        - Shapiro-Wilk Test: Uji formal terhadap normalitas residual. Dalam hasil penelitian, nilai p yang diperoleh mendekati 0.05, menunjukkan bahwa asumsi normalitas masih dapat diterima.
        - Kolmogorov-Smirnov Test: Uji ini membandingkan distribusi kumulatif residual dengan distribusi normal. Hasil uji menunjukkan bahwa tidak terdapat penyimpangan signifikan dari distribusi normal.
        - Uji Liliefors: Tidak digunakan secara eksplisit dalam penelitian ini, namun secara konsep serupa dengan Kolmogorov–Smirnov dan berguna ketika parameter distribusi tidak diketahui secara pasti.
    b. Uji Homoskedastisitas
        - Breusch-Pagan Test: Digunakan untuk menguji apakah varians residual tergantung pada nilai prediktor. Dalam penelitian ini, hasil uji menunjukkan nilai $p < 0.05$, yang mengindikasikan kemungkinan adanya heteroskedastisitas ringan.

- Modified Levene Test: Tidak digunakan secara eksplisit, namun prinsipnya serupa untuk menguji kesetaraan varians antar kelompok.

c. Uji Independensi

- Uji Durbin-Watson: Uji ini mendeteksi autokorelasi pada residual. Nilai DW mendekati 2, yang menunjukkan bahwa residual tidak mengalami autokorelasi dan asumsi independensi residual terpenuhi.

Secara keseluruhan, hasil analisis residual mendukung validitas model regresi akhir. Meskipun terdapat indikasi ringan terhadap heteroskedastisitas dan deviasi dari normalitas pada ekor distribusi, model tetap dapat digunakan untuk tujuan prediksi dan interpretasi karena penyimpangan tersebut tidak signifikan secara substantif.

2.6 Transformasi Model

Transformasi model dalam regresi linear digunakan untuk mengatasi berbagai permasalahan yang dapat mengganggu validitas dan performa model. Tujuan utama dari transformasi adalah untuk memenuhi asumsi regresi linear seperti linearitas hubungan, normalitas residual, homoskedastisitas, serta mengurangi pengaruh outlier dan multikolinearitas.

Dalam penelitian ini, transformasi dilakukan dengan mempertimbangkan beberapa pendekatan. Beberapa alasan utama untuk melakukan transformasi model meliputi:

1. Mengatasi Non-linearitas: Ketika hubungan antara prediktor dan respon bersifat melengkung atau tidak linier, transformasi digunakan untuk memperbaiki pola hubungan agar lebih mendekati linier.
2. Menstabilkan Varians (Homoskedastisitas): Residual yang tidak memiliki varians konstan dapat dikoreksi melalui transformasi.
3. Memperbaiki Normalitas Residual: Transformasi dapat digunakan untuk mendekatkan distribusi residual ke distribusi normal.
4. Mengatasi Outliers: Beberapa transformasi dapat mereduksi dampak data ekstrem terhadap hasil regresi.

Jenis Transformasi yang Sering Digunakan

1. Transformasi Kuadrat: Digunakan untuk menangkap hubungan non-linier, seperti pada variabel *Age*, *Stress Level*, dan *Heart Rate*. Variabel-variabel ini dimasukkan dalam bentuk I(Age^2), I(Stress Level^2), dan I(Heart Rate^2).

2. Transformasi Interaksi: Diterapkan untuk mengeksplorasi efek bersama dua variabel prediktor terhadap respon. Misalnya *Age* ; *Stress Level* dan *Physical Activity Level* ; *Stress Level* yang dimasukkan dalam model akhir karena memberikan kontribusi signifikan.
3. Transformasi Logaritmik dan Akar Kuadrat: Telah dipertimbangkan untuk mengatasi heteroskedastisitas dan ketidaksesuaian distribusi residual menggunakan transformasi logaritmik dan akar kuadrat, namun berdasarkan evaluasi grafik dan uji statistik yang didapatkan, transformasi ini tidak secara signifikan meningkatkan model, sehingga tidak digunakan pada model akhir.
4. Transformasi Box-Cox: Model juga dievaluasi menggunakan transformasi Box-Cox untuk mencari bentuk transformasi optimal berdasarkan nilai lambda (λ), namun hasil eksplorasi menunjukkan bahwa bentuk transformasi kuadrat dan interaksi sudah cukup efektif dalam memperbaiki asumsi-asumsi model.

Transformasi-transformasi tersebut secara keseluruhan berhasil meningkatkan kinerja model regresi dalam penelitian ini, yang ditunjukkan dengan membaiknya hasil uji asumsi dan peningkatan validitas prediksi model pada tahap validasi silang. Oleh karena itu, transformasi yang digunakan telah memenuhi tujuan untuk memperbaiki struktur model tanpa menambah kompleksitas yang tidak perlu.

2.6 Model Interaksi

Model interaksi dalam regresi linear digunakan untuk mengevaluasi apakah efek suatu variabel prediktor terhadap variabel respon berubah tergantung pada tingkat variabel prediktor lainnya. Konsep ini penting ketika pengaruh satu variabel terhadap respon tidak bersifat tetap, melainkan tergantung pada konteks atau nilai dari variabel lain.

1. Model regresi linear dengan interaksi dapat ditulis sebagai $Y= \beta0 + \beta1X1 + \beta2X2 + \beta3(X1 \times X2) + \epsilon$
   - $\beta0$: *Intercept*, yaitu nilai rata-rata respon (Y) ketika X1 dan X2 bernilai nol.
   - $\beta1$: Koefisien efek utama X1, menunjukkan perubahan rata-rata Y untuk setiap unit perubahan X1 saat X2 tetap konstan.
   - $\beta2$: Koefisien efek utama X2, menunjukkan perubahan rata-rata Y untuk setiap unit perubahan X2 saat X1 tetap konstan.
   - $\beta3$: Koefisien interaksi, menunjukkan bagaimana hubungan antara X1 dan Y berubah untuk setiap unit perubahan X2.

- **ε**: *Error term*, menggambarkan variasi dalam Y yang tidak dapat dijelaskan oleh model.

2. Langkah-langkah untuk Membangun Model Interaksi
   - Mempersiapkan Data: Dataset *Sleep Health and Lifestyle* yang digunakan telah diproses untuk menghapus duplikat, *outlier*, dan *missing values*.
   - Membangun Model dengan Interaksi: Variabel interaksi dibangun dari kombinasi variabel kuantitatif yang dianggap memiliki kemungkinan hubungan kompleks. Dalam penelitian ini, interaksi antara *Age* dengan *Stress Level* serta antara *Physical Activity Level* dengan *Stress Level* disertakan karena secara teoritis keduanya dapat berpengaruh secara bersama-sama terhadap durasi tidur.
   - Evaluasi dan Interpretasi: Koefisien interaksi diuji melalui model regresi linear dan diseleksi menggunakan metode *stepwise* berdasarkan nilai AIC. Hasil menunjukkan bahwa koefisien dari *Age*: *Stress Level* dan *Physical Activity Level:Stress Level* signifikan secara statistik, sehingga dimasukkan dalam model akhir.
3. Interpretasi Koefisien Interaksi
   - Jika koefisien interaksi β3 signifikan secara statistik, maka efek dari X1 terhadap Y akan berubah tergantung pada nilai X2. Misalnya, efek usia terhadap durasi tidur bisa berbeda tergantung tingkat stres individu.
   - Koefisien interaksi yang positif menunjukkan bahwa peningkatan X2 memperkuat pengaruh X1 terhadap Y, sedangkan koefisien negatif menunjukkan bahwa peningkatan X2 justru mengurangi pengaruh X1 terhadap Y.
4. Penerapan dalam Penelitian Ini

   Dalam model akhir, dua istilah interaksi yang signifikan adalah:
   - *Age:Stress Level*
   - *Physical Activity Level:Stress Level*

Kedua interaksi ini menggambarkan bahwa pengaruh usia dan aktivitas fisik terhadap durasi tidur sangat dipengaruhi oleh tingkat stres yang dialami responden. Hal ini mencerminkan kompleksitas dalam perilaku tidur yang tidak bisa dijelaskan hanya dengan efek langsung masing-masing variabel, melainkan juga perlu mempertimbangkan hubungan antar faktor.

Dengan memasukkan istilah interaksi yang tepat, model regresi menjadi lebih sensitif terhadap variasi dalam data dan mampu menangkap dinamika yang lebih realistis dalam hubungan antar variabel prediktor dan respon.

BAB III

METODE PENELITIAN

3.1 Deskrispsi Dataset

Dataset yang digunakan dalam penelitian ini berjudul "Sleep Health and Lifestyle Dataset" dan terdiri atas 374 observasi yang merepresentasikan individu dari berbagai latar belakang pekerjaan dan gaya hidup. Dataset ini berisi 13 variabel, baik numerik maupun kategorik, seperti *Person ID*, *Gender*, *Age*, *Occupation*, *Sleep Duration*, *Quality of Sleep*, *Physical Activity Level*, *Stress Level*, *BMI Category*, *Blood Pressure*, *Heart Rate*, *Daily Steps*, dan *Sleep Disorder*. Informasi ini mencakup data medis dan kebiasaan harian yang berpotensi memengaruhi kualitas serta durasi tidur, termasuk kondisi gangguan tidur seperti *insomnia* dan *sleep apnea*. Keberagaman atribut dalam dataset memungkinkan dilakukannya analisis komprehensif mengenai faktor-faktor yang berperan dalam pola tidur individu.

Dikarenakan penggunaan metode regresi linear, maka penting untuk menentukan variabel respon (dependen) dan variabel prediktor (independen). Variabel

respon adalah variabel yang nilainya akan diprediksi oleh model. Pada penelitian ini, durasi tidur (*Sleep Duration*) dipilih sebagai variabel respon karena merupakan aspek utama yang ingin dianalisis dan dipengaruhi oleh berbagai faktor gaya hidup. Sedangkan variabel prediktor adalah variabel-variabel yang diduga memiliki pengaruh terhadap durasi tidur. Dalam hal ini, variabel prediktor yang digunakan meliputi usia (*Age*), tingkat aktivitas fisik harian (*Physical Activity Level*), tingkat stres (*Stress Level*), detak jantung (*Heart Rate*), dan kualitas tidur (*Quality of Sleep*). Pemilihan variabel-variabel tersebut didasarkan pada relevansi teoritis serta dukungan dari data yang telah melalui tahap *pre-processing*, sehingga memungkinkan dilakukan analisis hubungan linear yang informatif dan bermakna. Dengan struktur data yang telah diproses dan ditentukan perannya, model regresi linear dapat dibangun untuk mengkaji bagaimana faktor-faktor tersebut saling berkontribusi terhadap variabilitas dalam durasi tidur individu.

3.2 Teknik Regresi Linear

Dalam penelitian ini, teknik analisis data yang digunakan adalah regresi linear berganda. Pendekatan ini dipilih untuk mengkaji pengaruh beberapa variabel bebas terhadap variabel terikat yaitu Sleep Duration, sekaligus membangun model prediksi yang akurat dan dapat dipercaya. Secara garis besar, proses pembentukan model regresi linear ini dilakukan melalui beberapa tahapan, yaitu sebagai berikut:

1. Exploratory Data Analysis (EDA)

   Tahap awal dilakukan dengan menganalisis data secara eksploratori guna memahami karakteristik data, mengidentifikasi data yang hilang, nilai pencilan (outlier), serta memastikan kesesuaian tipe data. Selain itu, distribusi data pada setiap variabel juga diperiksa untuk mengetahui potensi noise atau pola tertentu dalam data.

2. Pembentukan Model Awal

   Setelah data dinyatakan layak, model regresi linear awal dikembangkan dengan memasukkan semua variabel independen yang relevan. Dalam proses ini, selain komponen linier, dimasukkan pula unsur polinomial kuadrat untuk menangkap hubungan non-linear, serta interaksi antar variabel yang diduga memiliki efek bersama terhadap variabel Sleep Duration. Model awal kemudian disederhanakan menggunakan metode stepwise (both direction) guna memilih kombinasi variabel yang paling signifikan dan efisien secara statistik.

3. Pemeriksaan Multikolinearitas

   Uji multikolinearitas dilakukan untuk memastikan bahwa tidak terjadi korelasi tinggi antar variabel prediktor. Pemeriksaan dilakukan dengan menghitung nilai Variance Inflation Factor (VIF). Jika ditemukan variabel dengan nilai VIF yang melebihi batas wajar (umumnya >10), maka variabel tersebut dipertimbangkan untuk dieliminasi dari model agar menghindari bias estimasi dan menjaga kestabilan parameter.

4. Pengujian Normalitas Residual

Pengujian ini bertujuan untuk memastikan bahwa residual dari model terdistribusi normal. Pemeriksaan dilakukan dengan uji Shapiro-Wilk dan Kolmogorov–Smirnov, serta visualisasi seperti histogram dan Q-Q Plot. Jika residual tidak normal, maka dilakukan transformasi variabel dependen, salah satunya adalah transformasi reciprocal (1/y). Setelah transformasi, residual diuji kembali untuk memastikan asumsi normalitas terpenuhi.

5. Pengujian Homoskedastisitas

Pengujian homoskedastisitas dilakukan untuk melihat apakah residual memiliki ragam yang konstan pada seluruh nilai prediktor. Pemeriksaan ini dilakukan secara visual menggunakan plot residual terhadap nilai prediksi (fitted values) dan juga melalui uji statistik Breusch–Pagan. Jika ditemukan gejala heteroskedastisitas, dilakukan perbaikan model dengan menggunakan pendekatan Weighted Least Squares (WLS), di mana bobot dihitung berdasarkan model prediksi varians dari residual.

6. Pengujian Independensi Residual

Uji ini bertujuan untuk memastikan bahwa residual dari setiap observasi bersifat bebas atau tidak berkorelasi satu sama lain. Pengujian dilakukan menggunakan Durbin–Watson test, di mana nilai yang mendekati 2 menunjukkan tidak adanya autokorelasi antar residual.

7. Validasi Silang

Untuk mengukur kestabilan dan keandalan model secara menyeluruh, dilakukan validasi silang (cross-validation). Validasi ini menggunakan pendekatan k-fold cross validation dengan k=5. Hasil validasi ini berguna untuk menilai performa prediktif model pada data yang berbeda dari data pelatihan, sehingga estimasi kesalahan prediksi menjadi lebih objektif dan bebas dari bias overfitting.

3.3 Evaluasi Model

Untuk memastikan bahwa model regresi linear yang dibangun dapat digunakan secara akurat dan dapat diandalkan, maka dilakukan tahap evaluasi model. Evaluasi ini bertujuan untuk menilai sejauh mana asumsi-asumsi regresi terpenuhi, serta untuk mengukur performa prediksi model secara objektif. Evaluasi dilakukan melalui beberapa tahapan berikut:

1. Pengujian Normalitas Residual

   Pengujian ini dimaksudkan untuk mengetahui apakah residual dari model terdistribusi normal. Pemeriksaan dilakukan secara visual menggunakan histogram dan Q-Q Plot, serta dilengkapi dengan uji statistik Shapiro-Wilk dan Kolmogorov–Smirnov. Apabila hasil pengujian menunjukkan bahwa residual tidak normal, maka dilakukan transformasi terhadap variabel dependen, yaitu Sleep Duration, menjadi bentuk reciprocal (1/y) untuk memperbaiki distribusi residual. Setelah transformasi, normalitas residual diuji ulang untuk memastikan bahwa asumsi ini telah terpenuhi.

2. Pengujian Homoskedastisitas

   Pengujian homoskedastisitas bertujuan untuk melihat apakah residual memiliki varians yang konstan terhadap semua nilai prediktor (bersifat homoskedastik). Pengujian dilakukan melalui visualisasi plot residual terhadap nilai prediksi, serta secara statistik menggunakan uji Breusch–Pagan. Jika ditemukan masalah heteroskedastisitas, maka digunakan pendekatan Weighted Least Squares (WLS), di mana bobot dihitung berdasarkan model prediksi varians residual (diperoleh melalui regresi log residual kuadrat terhadap prediktor). Model WLS memberikan estimasi parameter yang lebih efisien dalam kondisi varians residual yang tidak homogen.

3. Pengujian Independensi Residual

   Uji independensi residual dilakukan untuk mengetahui apakah residual dari satu observasi saling bebas terhadap observasi lainnya. Pengujian dilakukan dengan Durbin–Watson test, di mana nilai yang mendekati 2 menunjukkan bahwa residual tidak mengalami autokorelasi. Jika asumsi ini terpenuhi, maka model dinyatakan bebas dari pengaruh hubungan antar residual.

4. Pengukuran Performa Prediksi

   Selain pengujian asumsi, evaluasi model juga mencakup penilaian terhadap performa prediktif model. Pengukuran ini dilakukan dengan melihat nilai-nilai statistik berikut:

   - Adjusted R-squared, untuk mengukur seberapa besar variasi dari variabel Sleep Duration yang dapat dijelaskan oleh model.
   - Root Mean Squared Error (RMSE) dan Mean Squared Error (MSE), untuk mengukur tingkat kesalahan prediksi. Semakin kecil nilai RMSE dan MSE, semakin baik performa model dalam memprediksi data.

- Penilaian RMSE dan MSE diperoleh melalui proses validasi silang (cross-validation) menggunakan k-fold (k = 5). Validasi silang ini dilakukan untuk menghindari bias terhadap data pelatihan dan untuk menguji ketepatan model pada data yang berbeda.

Dengan melakukan tahap-tahap evaluasi ini, diharapkan model regresi linear yang dibangun tidak hanya signifikan secara statistik, tetapi juga memenuhi asumsi-asumsi klasik regresi. Dengan demikian, model dapat digunakan secara valid dan hasil prediksi yang diperoleh dapat diandalkan.

BAB IV

HASIL DAN PEMBAHASAN

4.1 Strategi Model Building

Pada tahap ini, dilakukan penyusunan strategi untuk membangun model regresi linear yang optimal. Strategi yang diterapkan dalam penelitian ini menggunakan pendekatan *stepwise regression*, yaitu metode yang dilakukan secara bertahap dengan cara menambahkan atau mengeluarkan variabel prediktor berdasarkan tingkat signifikansi statistiknya sehingga hanya variabel yang berpengaruh yang dipertahankan dalam model.

Adapun tahapan dalam strategi *model building* ini mencakup langkah-langkah berikut:

4.1.1 Exploratory Data Analysis

Sebagai tahap awal dalam membangun model regresi linear, dilakukan *Exploratory Data Analysis* (EDA) untuk memahami pola dan karakteristik data, mendeteksi potensi masalah, serta memastikan data layak untuk dianalisis lebih lanjut. Berikut adalah langkah-langkah *preprocessing* yang telah diterapkan:

1. Penghapusan Variabel Index

   Langkah pertama dilakukan dengan menghapus variabel Person ID yang berfungsi hanya sebagai penanda baris (*index*), karena variabel ini tidak mengandung informasi yang relevan untuk dianalisis dalam model regresi.

2. Pemeriksaan Struktur Dataset

   Selanjutnya, dilakukan pengecekan struktur data untuk mengetahui jumlah baris dan variabel. Berdasarkan hasil pemeriksaan, data yang digunakan memiliki 374 baris dan 13 variabel, yang terdiri atas variabel numerik maupun kategorik yang relevan untuk dianalisis melalui regresi linear.

3. Deteksi dan Penanganan Duplikasi

   Dataset kemudian diperiksa adanya baris data yang terduplikasi. Data ganda dapat menyebabkan bias dengan memberi bobot berlebih pada

informasi tertentu. Hasil identifikasi menunjukkan terdapat 242 data duplikat, sehingga dibuat variabel baru sebagai penanda untuk mempermudah penghapusan. Setelah baris duplikat dihapus, dilakukan pengecekan ulang dan dipastikan tidak ada lagi data ganda.

4. Pemeriksaan Tipe Data

   Tiap variabel kemudian diverifikasi tipe datanya agar sesuai dengan jenis informasi yang diwakilinya. Variabel kategorik seperti *Gender*, *Occupation*, *BMI Category*, *Blood Pressure*, dan *Sleep Disorder* tercatat bertipe *object* atau *category*, sedangkan variabel numerik seperti *Age*, *Quality of Sleep*, *Physical Activity Level*, *Stress Level*, *Heart Rate*, dan *Daily Steps* bertipe numerik (*integer* atau *float*). Tidak ditemukan ketidaksesuaian sehingga perubahan tipe data tidak diperlukan.

5. Pemeriksaan dan Penanganan Noise

   Untuk meminimalkan adanya data yang bersifat *noise*, dilakukan pengecekan distribusi:

   - Pada variabel kategorik, frekuensi masing-masing kategori dianalisis melalui tabel distribusi. Hasilnya menunjukkan tidak terdapat kategori yang proporsinya terlalu kecil atau tidak wajar.
   - Pada variabel numerik, statistik deskriptif (nilai minimum, maksimum, rata-rata, dan standar deviasi) dihitung untuk memastikan rentang nilai masih masuk akal. Dari hasilnya dapat disimpulkan bahwa data dapat dipercaya untuk dianalisis lebih lanjut.

6. Identifikasi Outlier

   Proses deteksi *outlier* dilakukan pada variabel numerik seperti *Age*, *Sleep Duration*, *Quality of Sleep*, *Physical Activity Level*, *Stress Level*, *Heart Rate*, dan *Daily Steps* dengan metode *Z-Score* ($|Z| > 3$). Berdasarkan hasil pemeriksaan, ditemukan 1 outlier pada variabel *Heart Rate*. Oleh karena itu, dibuat variabel baru untuk membantu menghapus data outlier tersebut sehingga distribusi data tetap mendekati normal dan asumsi regresi dapat terpenuhi.

7. Penanganan Missing Values

   Data kemudian diperiksa terkait keberadaan nilai kosong. Ditemukan adanya 73 nilai kosong pada variabel *Sleep Disorder* yang

bertipe kategorik. Nilai yang hilang diimputasi menggunakan nilai modus agar tidak ada baris data yang terbuang dan jumlah observasi tetap optimal.

8. Analisis Korelasi

   Sebagai bagian dari eksplorasi lebih lanjut, dilakukan pengecekan korelasi antar variabel numerik (*Age*, *Sleep Duration*, *Quality of Sleep*, *Physical Activity Level*, *Stress Level*, *Heart Rate*, dan *Daily Steps*). Visualisasi matriks korelasi (*heatmap*) juga disusun untuk mempermudah interpretasi pola hubungan antar variabel.

9. Penyimpanan Dataset Final

   Setelah tahap pembersihan selesai, data hasil *preprocessing* disimpan dalam format file CSV di direktori *Google Colab*, lalu diunduh ke komputer untuk digunakan pada tahap analisis regresi berikutnya.

Secara keseluruhan, rangkaian EDA menunjukkan bahwa dataset telah dipersiapkan dengan matang. Data duplikat sudah dihapus, *outlier* telah ditangani, *missing values* telah diimputasi, tipe data sudah benar, dan distribusi data telah diperiksa untuk meminimalkan *noise*. Dengan demikian, dataset diyakini memenuhi syarat untuk digunakan dalam pembangunan model regresi linear yang valid dan dapat diandalkan.

4.1.2 Membangun Model Awal

Pada tahap ini, disusun sebuah model regresi linear awal yang mencakup semua variabel independen yang telah diidentifikasi relevan berdasarkan hasil *Exploratory Data Analysis* sebelumnya.

```
# Membuat model awal
model_awal <- lm(Sleep.Duration ~ Age + Quality.of.Sleep + Physical.Activity.Level
                 + Stress.Level + Heart.Rate + I(Age^2) + I(Quality.of.Sleep^2)
                 + I(Physical.Activity.Level^2) + I(Stress.Level^2) + I(Heart.Rate^2)
                 + Age:Quality.of.Sleep + Age:Physical.Activity.Level + Age:Stress.Level
                 + Age:Heart.Rate + Quality.of.Sleep:Physical.Activity.Level
                 + Quality.of.Sleep:Stress.Level + Quality.of.Sleep:Heart.Rate
                 + Physical.Activity.Level:Stress.Level
                 + Physical.Activity.Level:Heart.Rate + Stress.Level:Heart.Rate
                 + Gender + BMI.Category, data = data)
```

Gambar 1: Model Awal

Penyusunan model awal ini dilakukan dengan mempertimbangkan beberapa hal penting, yaitu:

- Variabel prediktor utama yang secara teoritis diyakini berpengaruh terhadap variabel respon
- Penambahan unsur polinomial (*kuadrat*) guna menangkap kemungkinan hubungan non-linier antara variabel independen dengan variabel dependen
- Interaksi antar variabel independen yang secara teori maupun temuan empiris berpotensi memengaruhi hubungan dengan variabel dependen.

Dalam penelitian ini, variabel dependen yang dianalisis adalah *Sleep Duration*, sedangkan variabel bebasnya meliputi:

- Variabel numerik, antara lain *Age*, *Quality of Sleep*, *Physical Activity Level*, *Stress Level*, dan *Heart Rate*.
- Interaksi di antara variabel numerik, misalnya interaksi *Age* dengan *Quality of Sleep*, *Age* dengan *Physical Activity Level*, serta kombinasi interaksi lainnya.
- Komponen polinomial berupa pangkat dua pada masing-masing variabel numerik untuk menggambarkan kemungkinan pola hubungan yang melengkung.
- Ditambah variabel kategorik seperti *Gender* dan *BMI Category* yang juga dimasukkan sebagai prediktor dalam model awal.

4.1.3 Membangun Model Formula Stepwise

Dalam upaya mendapatkan model regresi linear yang optimal, dilakukan pemilihan variabel menggunakan metode stepwise. Metode ini berfungsi untuk menyaring variabel independen yang paling berpengaruh terhadap variabel dependen. Pemilihan variabel dilakukan dengan bantuan fungsi *step()* pada software R, dengan opsi arah seleksi both, yang berarti pemilihan dilakukan secara maju (forward) dan mundur (backward) secara simultan. Berikut ditampilkan cuplikan kode R yang digunakan:

```
# Membangun formula Stepwise
stepwise_model <- step(model_awal, direction = "both", trace = TRUE)
summary(stepwise_model)
formula(stepwise_model)
```

Gambar 2: Code dari *stepwise_model*

Output dari kode tersebut memberikan formula akhir dari model yang dipilih secara otomatis oleh proses stepwise. Model akhir tersebut memuat variabel utama, kuadrat dari beberapa variabel (untuk mendeteksi hubungan

non-linear), serta interaksi antara beberapa pasangan variabel. Berikut adalah hasil formula akhir model stepwise:

```
> formula(stepwise_model)
Sleep.Duration ~ Age + Quality.of.Sleep + Physical.Activity.Level +
    Stress.Level + Heart.Rate + I(Age^2) + I(Quality.of.Sleep^2) +
    I(Physical.Activity.Level^2) + I(Stress.Level^2) + I(Heart.Rate^2) +
    Gender + BMI.Category + Age:Quality.of.Sleep + Age:Physical.Activity.Level +
    Age:Heart.Rate + Quality.of.Sleep:Physical.Activity.Level +
    Quality.of.Sleep:Stress.Level + Quality.of.Sleep:Heart.Rate +
    Physical.Activity.Level:Heart.Rate + Stress.Level:Heart.Rate
```

Gambar 3: Output dari *formula(stepwise_model)*

4.1.4 Uji Multikolinearitas

Pemeriksaan multikolinearitas dilakukan dengan cara menghitung nilai *Variance Inflation Factor (VIF)* pada setiap variabel, termasuk variabel interaksi dan unsur polinomial. Secara umum, variabel dengan nilai VIF melebihi 10 atau *Generalized VIF (GVIF)* yang besar dapat mengindikasikan adanya masalah multikolinearitas yang perlu ditangani. Oleh karena itu, dilakukan proses seleksi model secara bertahap untuk meminimalkan gejala multikolinearitas dan meningkatkan stabilitas model.

Proses seleksi dilakukan melalui beberapa tahap dengan kriteria sebagai berikut:

- Tahap Pertama

```
> vif(model_reduced)
there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif
                                              GVIF Df GVIF^(1/(2*Df))
Age                                     3194.021658  1       56.515676
Quality.of.Sleep                       20841.587595  1      144.366158
Physical.Activity.Level                15133.458691  1      123.018123
Stress.Level                            8558.822966  1       92.513907
Heart.Rate                             11278.101805  1      106.198408
I(Age^2)                                 980.526866  1       31.313366
I(Quality.of.Sleep^2)                   5412.781691  1       73.571609
I(Physical.Activity.Level^2)             186.646932  1       13.661879
I(Stress.Level^2)                       2557.847641  1       50.575168
I(Heart.Rate^2)                         7628.228395  1       87.339730
Gender                                     3.003163  1        1.732964
BMI.Category                              78.310690  3        2.068411
Age:Quality.of.Sleep                    1964.889309  1       44.327072
Age:Physical.Activity.Level              175.476510  1       13.246755
Age:Heart.Rate                          1892.007356  1       43.497211
Quality.of.Sleep:Physical.Activity.Level 1530.036073  1       39.115676
Quality.of.Sleep:Stress.Level           1354.990419  1       36.810195
Quality.of.Sleep:Heart.Rate             6125.725597  1       78.267015
Physical.Activity.Level:Heart.Rate     10206.856275  1      101.028987
Stress.Level:Heart.Rate                10056.876012  1      100.283977
```

Gambar 4: Output dari *vif(model_reduced)*

Pada tahap pertama, dilakukan peninjauan terhadap hasil perhitungan VIF pada model awal. Batas nilai GVIF^(1/(2*Df)) yang digunakan untuk seleksi adalah maksimal 100. Variabel-variabel yang

memiliki nilai di atas ambang ini dianggap berpotensi menimbulkan multikolinearitas yang sangat tinggi di dalam model. Oleh sebab itu, variabel-variabel tersebut diputuskan untuk dieliminasi agar korelasi antar prediktor dapat ditekan dan estimasi parameter model menjadi lebih akurat serta stabil.

- Tahap Kedua

Setelah variabel-variabel dengan GVIF^(1/(2Df)) melebihi 100 dihapus, model disusun ulang tanpa variabel tersebut. Kemudian dilakukan penghitungan VIF kembali pada model kedua untuk memastikan apakah masih ada prediktor yang menunjukkan indikasi multikolinearitas tinggi. Pada tahap kedua ini, kriteria seleksi difokuskan pada variabel dengan nilai GVIF^(1/(2Df)) di atas 50. Variabel-variabel yang melebihi ambang batas ini selanjutnya dievaluasi lebih lanjut dan dihilangkan dari model apabila dinilai tidak memiliki kontribusi yang penting, baik secara statistik maupun dari segi substansi teori.

```
# Seleksi variabel
stepwise_model1 <- update(model_awal, . ~ Age + Stress.Level + I(Age^2) +
                          I(Quality.of.Sleep^2) + I(Physical.Activity.Level^2) +
                          I(Stress.Level^2) + I(Heart.Rate^2) + Gender + BMI.Category
                          Age:Quality.of.Sleep + Age:Physical.Activity.Level +
                          Age:Heart.Rate + Quality.of.Sleep:Physical.Activity.Level +
                          Quality.of.Sleep:Stress.Level + Quality.of.Sleep:Heart.Rate
                          Stress.Level:Heart.Rate, data = data)
summary(stepwise_model1)
vif(stepwise_model1)
```

Gambar 5: Code dari *stepwise_model1*

```
> vif(stepwise_model1)

there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif

                                            GVIF Df GVIF^(1/(2*Df))
Age                                 3033.152522  1       55.074064
Stress.Level                        3797.523590  1       61.624030
I(Age^2)                             517.319697  1       22.744663
I(Quality.of.Sleep^2)                662.064334  1       25.730611
I(Physical.Activity.Level^2)          54.760547  1        7.400037
I(Stress.Level^2)                   1239.634656  1       35.208446
I(Heart.Rate^2)                      837.232590  1       28.934972
Gender                                 2.720777  1        1.649478
BMI.Category                          34.024520  3        1.800108
Age:Quality.of.Sleep                1217.054555  1       34.886309
Age:Physical.Activity.Level          154.597526  1       12.433725
Age:Heart.Rate                      1754.618262  1       41.888164
Quality.of.Sleep:Physical.Activity.Level  143.335595  1       11.972284
Stress.Level:Quality.of.Sleep        555.703357  1       23.573361
Quality.of.Sleep:Heart.Rate         1071.023859  1       32.726501
Stress.Level:Heart.Rate             4616.438041  1       67.944375
```

Gambar 6: Output dari *vif(model_1)*

- Tahap Ketiga

Pada tahap ketiga, penyusunan model dilakukan kembali dengan mengacu pada hasil seleksi pada tahap kedua. Penghitungan VIF diulang

sekali lagi untuk memverifikasi apakah masih terdapat variabel yang berpotensi menimbulkan multikolinearitas tinggi. Pada tahap ini dilakukan penyeleksian secara bertahap dalam penurunan nilai VIF. Pertama, seleksi variabel Age:Quality of Sleep dengan nilai GVIF^(1/(2Df)) sebesar 25.610557. Kedua, seleksi variabel I(Quality of Sleep^2) sebesar 12.665489. Terakhir, seleksi variabel Age:Physical Activity Level Variabel dengan nilai GVIF^(1/(2Df)) sebesar 10.890526. Dengan demikian, model regresi linear akhir diharapkan lebih stabil, minim multikolinearitas, serta dapat memberikan hasil interpretasi yang lebih dapat diandalkan.

```
stepwise_model2 <- update(model_awal, . ~ I(Age^2) + I(Quality.of.Sleep^2) +
                            I(Physical.Activity.Level^2) + I(Stress.Level^2) +
                            I(Heart.Rate^2) + Gender + BMI.Category +
                            Age:Quality.of.Sleep + Age:Physical.Activity.Level +
                            Age:Heart.Rate + Quality.of.Sleep:Physical.Activity.Level +
                            Quality.of.Sleep:Stress.Level + Quality.of.Sleep:Heart.Rate,
                          data = data)
summary(stepwise_model2)
vif(stepwise_model2)
```

Gambar 7: Code dari *stepwise_model2*

```
> vif(stepwise_model2)

there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif

                                             GVIF Df GVIF^(1/(2*Df))
I(Age^2)                                510.636853  1       22.597275
I(Quality.of.Sleep^2)                   296.998773  1       17.233652
I(Physical.Activity.Level^2)             45.271676  1        6.728423
I(Stress.Level^2)                        69.028089  1        8.308314
I(Heart.Rate^2)                          31.258735  1        5.590951
Gender                                    2.460216  1        1.568508
BMI.Category                             16.556317  3        1.596469
Age:Quality.of.Sleep                    655.900619  1       25.610557
Age:Physical.Activity.Level             137.244684  1       11.715148
Age:Heart.Rate                           97.605012  1        9.879525
Quality.of.Sleep:Physical.Activity.Level 108.983320  1       10.439508
Quality.of.Sleep:Stress.Level            30.263209  1        5.501201
Quality.of.Sleep:Heart.Rate             116.949438  1       10.814316
```

Gambar 8: Output dari *vif(stepwise_model2)*

```
stepwise_model3 <- update(model_awal, . ~ I(Age^2) + I(Quality.of.Sleep^2) +
                            I(Physical.Activity.Level^2) + I(Stress.Level^2) +
                            I(Heart.Rate^2) + Gender + BMI.Category +
                            Age:Physical.Activity.Level + Age:Heart.Rate +
                            Quality.of.Sleep:Physical.Activity.Level +
                            Quality.of.Sleep:Stress.Level +
                            Quality.of.Sleep:Heart.Rate,
                          data = data)
summary(stepwise_model3)
vif(stepwise_model3)
```

Gambar 9: Code dari *stepwise_model3*

```
> vif(stepwise_model3)

there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif

                                              GVIF Df GVIF^(1/(2*Df))
I(Age^2)                                105.990572  1       10.295172
I(Quality.of.Sleep^2)                   160.414615  1       12.665489
I(Physical.Activity.Level^2)             43.022252  1        6.559135
I(Stress.Level^2)                        65.429009  1        8.088820
I(Heart.Rate^2)                          29.305912  1        5.413494
Gender                                    2.347044  1        1.532007
BMI.Category                             15.187675  3        1.573676
Age:Physical.Activity.Level             136.621881  1       11.688536
Age:Heart.Rate                           92.131442  1        9.598512
Physical.Activity.Level:Quality.of.Sleep 106.257367  1       10.308121
Quality.of.Sleep:Stress.Level            29.672886  1        5.447282
Heart.Rate:Quality.of.Sleep             116.645368  1       10.800249
```

Gambar 10: Output dari *vif(stepwise_model3)*

```
stepwise_model4 <- update(model_awal, . ~ I(Age^2) + I(Physical.Activity.Level^2) +
                            I(Stress.Level^2) + I(Heart.Rate^2) + Gender + BMI.Category +
                            Age:Physical.Activity.Level + Age:Heart.Rate +
                            Quality.of.Sleep:Physical.Activity.Level +
                            Quality.of.Sleep:Stress.Level +
                            Quality.of.Sleep:Heart.Rate,
                          data = data)
summary(stepwise_model4)
vif(stepwise_model4)
```

Gambar 11: Code dari *stepwise_model4*

```
> vif(stepwise_model4)

there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif

                                              GVIF Df GVIF^(1/(2*Df))
I(Age^2)                                 92.440598  1        9.614603
I(Physical.Activity.Level^2)             42.928749  1        6.552003
I(Stress.Level^2)                        49.487974  1        7.034769
I(Heart.Rate^2)                          16.471851  1        4.058553
Gender                                    2.344697  1        1.531240
BMI.Category                             13.448535  3        1.542100
Age:Physical.Activity.Level             118.603558  1       10.890526
Age:Heart.Rate                           91.162562  1        9.547909
Physical.Activity.Level:Quality.of.Sleep 93.123934  1        9.650074
Quality.of.Sleep:Stress.Level            17.631845  1        4.199029
Heart.Rate:Quality.of.Sleep              21.230445  1        4.607651
```

Gambar 12: Output dari *vif(stepwise_model4)*

- Tahap Keempat

Pada tahap keempat, model disusun kembali dengan mempertimbangkan hasil seleksi variabel pada tahap sebelumnya. Pada tahap ini, dilakukan pemeriksaan ulang terhadap nilai GVIF^(1/(2*Df)) untuk setiap prediktor dalam model terbaru. Langkah ini bertujuan memastikan bahwa semua variabel yang dipertahankan telah memiliki nilai GVIF yang berada di bawah ambang batas 10. Dengan demikian, model regresi linear terbaru diharapkan benar-benar minim multikolinearitas sehingga estimasi koefisiennya lebih stabil dan interpretasi hasilnya lebih dapat diandalkan.

```
stepwise_model5 <- update(model_awal, . ~ I(Age^2) + I(Physical.Activity.Level^2) +
                          I(Stress.Level^2) + I(Heart.Rate^2) + Gender + BMI.Category +
                          Age:Heart.Rate + Quality.of.Sleep:Physical.Activity.Level +
                          Quality.of.Sleep:Stress.Level +
                          Quality.of.Sleep:Heart.Rate,
                        data = data)
summary(stepwise_model5)
vif(stepwise_model5)
vif(stepwise_model5, type = 'predictor')
```

Gambar 13: Code dari *stepwise_model5*

```
> vif(stepwise_model5)
there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif
                                              GVIF Df GVIF^(1/(2*Df))
I(Age^2)                                 91.978044  1        9.590518
I(Physical.Activity.Level^2)             34.482378  1        5.872170
I(Stress.Level^2)                        41.720686  1        6.459155
I(Heart.Rate^2)                          15.761886  1        3.970124
Gender                                    2.158486  1        1.469179
BMI.Category                             11.445326  3        1.501198
Age:Heart.Rate                           79.120409  1        8.894965
Quality.of.Sleep:Physical.Activity.Level 48.166924  1        6.940239
Quality.of.Sleep:Stress.Level            16.645030  1        4.079832
Heart.Rate:Quality.of.Sleep              20.946507  1        4.576735
```

Gambar 14: Output dari *vif(stepwise_model5)*

- Tahap Kelima

Pada tahap kelima, disusun model regresi linear yang diperbarui dengan mendasarkan pada hasil pemilihan variabel dari tahap sebelumnya. Pada tahap ini, variabel-variabel prediktor diatur kembali menggunakan variabel yang terpilih melalui proses seleksi pada model ketiga. Sementara itu, variabel dependen yang dianalisis tetap difokuskan pada *Sleep Duration* sebagai variabel utama dalam penelitian ini.

```
model_terbaru <- lm(Sleep.Duration ~ I(Age^2) + I(Physical.Activity.Level^2) +
                    I(Stress.Level^2) + I(Heart.Rate^2) + Gender + BMI.Category +
                    Age:Heart.Rate + Quality.of.Sleep:Physical.Activity.Level +
                    Quality.of.Sleep:Stress.Level +
                    Quality.of.Sleep:Heart.Rate, data = data)
summary(model_terbaru)
vif(model_terbaru)
```

Gambar 15: Code dari *model_terbaru*

Pembangunan model ini dilakukan dengan menjalankan fungsi *lm()* pada perangkat lunak R. Selanjutnya, perintah *summary(model_terbaru)* digunakan untuk melihat ringkasan hasil estimasi koefisien, uji signifikansi parameter, serta *vif(model_terbaru)*

dijalankan untuk memeriksa nilai VIF guna memastikan bahwa model terbaru telah bebas dari masalah multikolinearitas yang berlebihan.

```
> summary(model_terbaru)

Call:
lm(formula = Sleep.Duration ~ I(Age^2) + I(Physical.Activity.Level^2) +
    I(Stress.Level^2) + I(Heart.Rate^2) + Gender + BMI.Category +
    Age:Heart.Rate + Quality.of.Sleep:Physical.Activity.Level +
    Quality.of.Sleep:Stress.Level + Quality.of.Sleep:Heart.Rate,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.97290 -0.17618 -0.05442  0.15553  0.71497

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                            5.625e+00  4.941e-01  11.385  < 2e-16 ***
I(Age^2)                               1.746e-03  3.045e-04   5.735 7.65e-08 ***
I(Physical.Activity.Level^2)           2.026e-04  5.712e-05   3.546 0.000561 ***
I(Stress.Level^2)                     -3.751e-02  8.020e-03  -4.677 7.80e-06 ***
I(Heart.Rate^2)                        6.255e-04  1.377e-04   4.541 1.36e-05 ***
GenderMale                             4.280e-01  7.026e-02   6.092 1.44e-08 ***
BMI.CategoryNormal Weight              5.923e-02  9.384e-02   0.631 0.529193
BMI.CategoryObese                     -7.670e-01  2.012e-01  -3.811 0.000221 ***
BMI.CategoryOverweight                -4.322e-01  9.903e-02  -4.364 2.75e-05 ***
Age:Heart.Rate                        -1.738e-03  3.591e-04  -4.839 3.98e-06 ***
Quality.of.Sleep:Physical.Activity.Level -2.838e-03  9.434e-04  -3.008 0.003216 **
Quality.of.Sleep:Stress.Level          1.426e-02  1.312e-02   1.087 0.279303
Heart.Rate:Quality.of.Sleep            2.766e-03  1.551e-03   1.783 0.077110 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2737 on 118 degrees of freedom
Multiple R-squared:  0.8863,    Adjusted R-squared:  0.8747
F-statistic: 76.63 on 12 and 118 DF,  p-value: < 2.2e-16
```

Gambar 16: Output dari *summary(model_terbaru)*

```
> vif(model_terbaru)
there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif
                                        GVIF Df GVIF^(1/(2*Df))
I(Age^2)                           91.978044  1        9.590518
I(Physical.Activity.Level^2)       34.482378  1        5.872170
I(Stress.Level^2)                  41.720686  1        6.459155
I(Heart.Rate^2)                    15.761886  1        3.970124
Gender                              2.158486  1        1.469179
BMI.Category                       11.445326  3        1.501198
Age:Heart.Rate                     79.120409  1        8.894965
Quality.of.Sleep:Physical.Activity.Level 48.166924  1        6.940239
Quality.of.Sleep:Stress.Level      16.645030  1        4.079832
Heart.Rate:Quality.of.Sleep        20.946507  1        4.576735
```

Gambar 17: Output dari *vif(model_terbaru)*

Berdasarkan hasil output *summary(model_terbaru)*, tampak bahwa masih ada beberapa variabel prediktor dengan nilai p-value yang melebihi 0,05. Hal ini menunjukkan bahwa variabel tersebut belum signifikan secara statistik pada tingkat signifikansi 5%. Oleh karena itu, variabel-variabel yang tidak signifikan ini perlu dipertimbangkan untuk dihapus, agar model regresi linear akhir hanya memuat prediktor yang memberikan pengaruh nyata terhadap variabel terikat.

Di sisi lain, hasil output dari *vif(model_terbaru)* memperlihatkan bahwa semua nilai GVIF^(1/(2*Df)) sudah berada di bawah angka 10. Artinya, model terbaru ini dapat dikatakan cukup bebas dari masalah

multikolinearitas yang tinggi, sehingga estimasi koefisien regresi menjadi lebih stabil dan hasil interpretasi dapat diandalkan.

Dengan kondisi ini, tahap berikutnya adalah melanjutkan proses seleksi dengan memfokuskan pada variabel-variabel yang nilai p-value nya masih melebihi batas signifikansi.

- Tahap Keenam

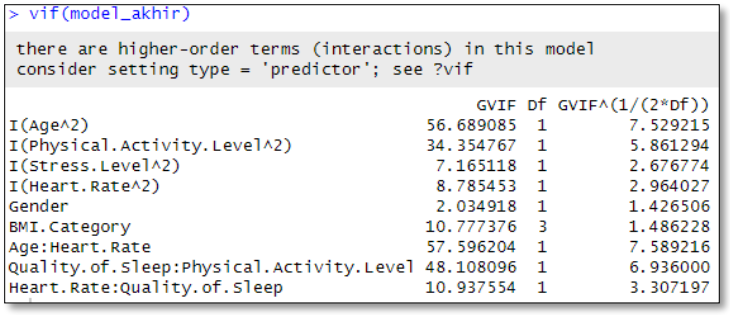
Pada tahap keenam, dilakukan proses seleksi lanjutan dengan cara menghapus variabel prediktor yang memiliki p-value di atas 0,05 berdasarkan hasil uji signifikansi pada model sebelumnya. Setelah variabel-variabel yang tidak signifikan ini dieliminasi, dilakukan regresi ulang untuk membentuk model akhir. Model akhir inilah yang kemudian dijadikan acuan dalam pengujian asumsi-asumsi regresi klasik berikutnya sekaligus sebagai dasar interpretasi hasil analisis.

```
model_akhir <- lm(Sleep.Duration ~ I(Age^2) + I(Physical.Activity.Level^2) +
                    I(Stress.Level^2) + I(Heart.Rate^2) + Gender + BMI.Category +
                    Age:Heart.Rate + Quality.of.Sleep:Physical.Activity.Level +
                    Heart.Rate:Quality.of.Sleep,
                  data = data)
summary(model_akhir)
vif(model_akhir)
```

Gambar 18: Code dari *model_akhir*

```
> summary(model_akhir)

Call:
lm(formula = Sleep.Duration ~ I(Age^2) + I(Physical.Activity.Level^2) +
    I(Stress.Level^2) + I(Heart.Rate^2) + Gender + BMI.Category +
    Age:Heart.Rate + Quality.of.Sleep:Physical.Activity.Level +
    Heart.Rate:Quality.of.Sleep, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.94429 -0.17371 -0.06026  0.16370  0.72086

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                            5.563e+00  4.911e-01  11.327  < 2e-16 ***
I(Age^2)                               1.541e-03  2.392e-04   6.443 2.60e-09 ***
I(Physical.Activity.Level^2)           2.064e-04  5.706e-05   3.616 0.000439 ***
I(Stress.Level^2)                     -2.958e-02  3.326e-03  -8.893 7.77e-15 ***
I(Heart.Rate^2)                        5.259e-04  1.029e-04   5.111 1.24e-06 ***
GenderMale                             4.463e-01  6.827e-02   6.537 1.64e-09 ***
BMI.CategoryNormal Weight              7.812e-02  9.229e-02   0.846 0.398997
BMI.CategoryObese                     -7.524e-01  2.010e-01  -3.744 0.000280 ***
BMI.CategoryOverweight                -4.084e-01  9.665e-02  -4.225 4.70e-05 ***
Age:Heart.Rate                        -1.534e-03  3.066e-04  -5.004 1.96e-06 ***
Quality.of.Sleep:Physical.Activity.Level -2.874e-03  9.436e-04  -3.045 0.002861 **
Heart.Rate:Quality.of.Sleep            3.932e-03  1.122e-03   3.505 0.000645 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2739 on 119 degrees of freedom
Multiple R-squared:  0.8851,    Adjusted R-squared:  0.8745
F-statistic: 83.36 on 11 and 119 DF,  p-value: < 2.2e-16
```

Gambar 19: Output dari *summary(model_akhir)*



```
> vif(model_akhir)

there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif

                                          GVIF Df GVIF^(1/(2*Df))
I(Age^2)                             56.689085  1        7.529215
I(Physical.Activity.Level^2)         34.354767  1        5.861294
I(Stress.Level^2)                     7.165118  1        2.676774
I(Heart.Rate^2)                       8.785453  1        2.964027
Gender                                2.034918  1        1.426506
BMI.Category                         10.777376  3        1.486228
Age:Heart.Rate                       57.596204  1        7.589216
Quality.of.Sleep:Physical.Activity.Level 48.108096  1        6.936000
Heart.Rate:Quality.of.Sleep          10.937554  1        3.307197
```

Gambar 20: Output dari *vif(model_akhir)*

Berdasarkan hasil *summary(model_akhir)*, dapat diketahui bahwa seluruh variabel prediktor yang tersisa dalam model memiliki nilai p-value kurang dari 0,05. Artinya, setiap variabel terbukti signifikan secara statistik pada tingkat signifikansi 5%, sehingga masing-masing prediktor diyakini memberikan pengaruh yang bermakna terhadap variabel dependen *Sleep Duration*.

Selain itu, dari hasil pemeriksaan multikolinearitas melalui *vif(model_akhir)*, terlihat bahwa semua nilai GVIF^(1/(2*Df)) sudah berada di bawah nilai ambang 10. Kondisi ini menunjukkan bahwa model akhir dapat dikatakan bebas dari masalah multikolinearitas yang tinggi, sehingga estimasi koefisien regresi dapat dihasilkan secara stabil dan interpretasi model menjadi lebih dapat dipercaya.

Dari hasil pemodelan akhir yang telah diperoleh, diperlihatkan bahwa terdapat sejumlah faktor yang memiliki pengaruh nyata terhadap variabel durasi tidur (**Sleep Duration**) dengan nilai p-value di bawah 0,05. Faktor-faktor tersebut antara lain:

- $I(Age^2)$ dengan p-value 2.60e-09.
- $I(Physical.Activity.Level^2)$ dengan p-value 0.000439.
- $I(Stress.Level^2)$ dengan p-value 7.77e-15.
- $I(Heart.Rate^2)$ dengan p-value 1.24e-06.
- Jenis kelamin (**Gender**) dengan p-value 1.64e-09.
- Kategori IMT (**BMI.Category**) dengan p-value 4.70e-05.
- Interaksi antara usia dan detak jantung (**Age:Heart.Rate**) dengan p-value 1.96e-06.

- Interaksi antara kualitas tidur dan aktivitas fisik (**Quality.of.Sleep:Physical.Activity.Level**) dengan p-value 0.002861.
- Interaksi antara detak jantung dan kualitas tidur (**Heart.Rate;Quality.of.Sleep**) dengan p-value 0.000645.

Dengan demikian, dapat disimpulkan bahwa ketujuh variabel tersebut merupakan faktor yang secara statistik berpengaruh signifikan dalam menjelaskan variasi durasi tidur responden.

4.2 Interpretasi Hasil Regresi Linear

1. Uji Normalitas Residual

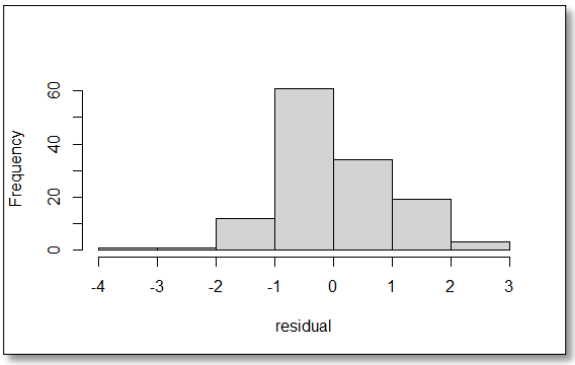
Pemeriksaan normalitas residual bertujuan untuk menilai apakah sisa (residual) dari model regresi linear mengikuti distribusi normal. Asumsi ini penting karena berkaitan langsung dengan validitas uji signifikansi parameter dan estimasi interval kepercayaan. Pengujian dilakukan melalui dua pendekatan, yaitu visualisasi dan analisis statistik.

```
# Uji Normalitas
resid1 <- residuals(model_akhir)
resid2 <- rstandard(model_akhir)
hist(resid2, xlab = "residual", main = "")
ols_plot_resid_qq(model_akhir)

shapiro.test(resid1)
ks.test(resid2, "pnorm", mean = 0, sd = 1)
ols_test_normality(model_akhir)
```

Gambar 21: Code Uji Normalitas

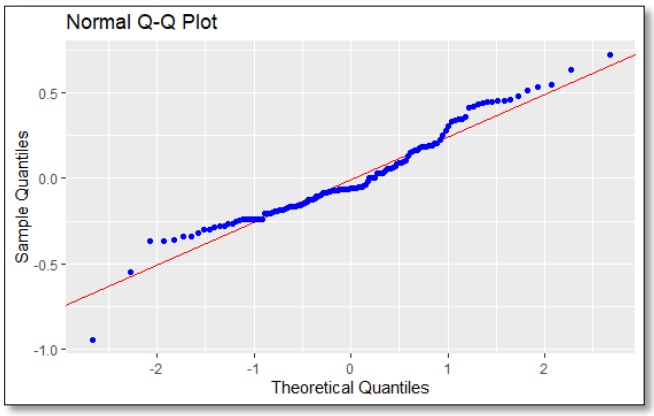
- Visualisasi Histogram dan Q-Q Plot

Gambar berikut menunjukkan histogram dan Q-Q plot dari residual standar:



Gambar 22: Output berupa Histogram

Gambar 16 memperlihatkan histogram residual standar yang menunjukkan pola mendekati distribusi normal, dengan puncak di sekitar nol. Terdapat sedikit kemiringan ke kiri, namun tidak signifikan dan tidak ditemukan outlier ekstrem. Secara umum, distribusi residual masih dapat dianggap normal.



Gambar 23: Output berupa Q-Q Plot

Pada Gambar 17, Q-Q plot menunjukkan bahwa sebagian besar titik mengikuti garis normal teoritis. Meskipun terdapat sedikit penyimpangan pada bagian ekor, pola tersebut masih dalam batas wajar. Oleh karena itu, asumsi normalitas residual dianggap terpenuhi.

- Uji Statistik

5. Uji Shapiro-Wilk

```
        Shapiro-Wilk normality test

data:  resid1
W = 0.95733, p-value = 0.0004091
```

Gambar 24: Output dari Uji Shapiro-Wilk

Hasil uji Shapiro-Wilk menunjukkan nilai statistik W = 0.95733 dengan p-value = 0.0004091. Karena p-value < 0.05, maka dapat disimpulkan bahwa residual tidak berdistribusi normal secara signifikan pada taraf kepercayaan 95%.

2. Uji Kolmogorov-Smirnov

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  resid2
D = 0.12844, p-value = 0.02654
alternative hypothesis: two-sided
```

Gambar 25: Output dari Uji Kolmogorov-Smirnov

Hasil uji Kolmogorov-Smirnov menunjukkan nilai D = 0.12844 dengan p-value = 0.02654. Sama seperti sebelumnya, p-value < 0.05, sehingga terdapat cukup bukti untuk menolak hipotesis nol sehingga residual tidak mengikuti distribusi normal secara signifikan.

Berdasarkan hasil uji normalitas residual sebelumnya, baik secara visual maupun statistik, ditemukan bahwa asumsi normalitas belum sepenuhnya terpenuhi. Meskipun histogram dan Q-Q plot menunjukkan distribusi residual yang relatif normal, hasil uji formal Shapiro-Wilk dan Kolmogorov-Smirnov memberikan p-value < 0.05, yang mengindikasikan bahwa residual tidak berdistribusi normal secara signifikan pada taraf kepercayaan 95%.

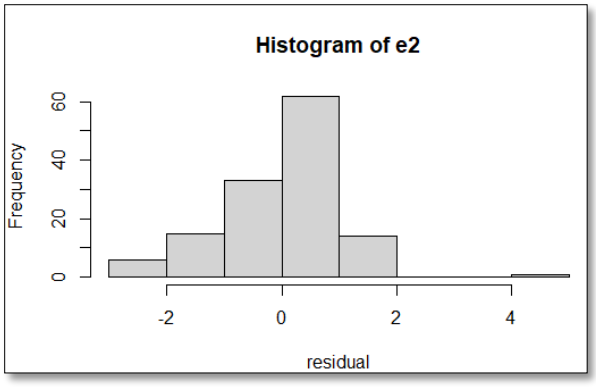
Untuk mengatasi pelanggaran asumsi normalitas itu, dilakukan transformasi terhadap variabel dependen **Sleep Duration**. Salah satu transformasi yang umum digunakan ketika residual tidak normal adalah transformasi **invers (1/y)**. Oleh karena itu, dibentuk model baru dengan bentuk transformasi 1/(*Sleep.Duration*), dengan struktur model seperti dibawah ini:

```
#Transformasi 1/y
model_1y <- lm(1/(Sleep.Duration) ~ I(Age^2) + I(Physical.Activity.Level^2) +
                I(Stress.Level^2) + I(Heart.Rate^2) + Gender + BMI.Category +
                Age:Heart.Rate + Quality.of.Sleep:Physical.Activity.Level +
                Heart.Rate:Quality.of.Sleep, data = data)
```

Gambar 26: Transformasi Invers $(\frac{1}{y})$

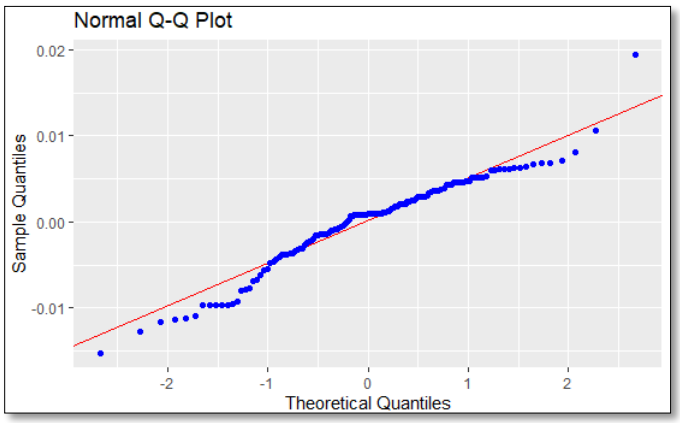
- Visualisasi Histogram dan Q-Q Plot

  Gambar berikut menunjukkan histogram dan Q-Q plot dari Transformasi Invers $(\frac{1}{y})$:



Gambar 27: Output berupa Histogram

Gambar 21 menunjukkan histogram residual standar dari model hasil transformasi 1/(*Sleep.Duration*). Distribusi residual tampak lebih simetris dengan puncak di sekitar nol dan sebaran yang relatif seimbang. Meski terdapat satu outlier di sisi kanan, secara keseluruhan transformasi ini berhasil memperbaiki distribusi residual sehingga lebih mendekati normal. Hal ini menunjukkan bahwa asumsi normalitas lebih terpenuhi dibandingkan model sebelumnya.



Gambar 28: Output berupa Q-Q Plot

Gambar 22 menampilkan Q-Q plot residual standar dari model hasil transformasi 1/(Sleep.Duration). Sebagian besar titik mengikuti garis diagonal, menunjukkan bahwa residual mendekati distribusi normal. Meski terdapat sedikit penyimpangan di kuantil ekstrem, pola tersebut masih dalam batas wajar. Dibandingkan model awal, Q-Q plot ini menunjukkan perbaikan, sehingga transformasi dinilai efektif dalam memenuhi asumsi normalitas.

- Uji Statistik

1. Uji Shapiro-Wilk

```
        Shapiro-Wilk normality test

data:  e2
W = 0.95395, p-value = 0.0002172
```

Gambar 29: Output dari Uji Shapiro-Wilk

Hasil uji Shapiro-Wilk menunjukkan nilai statistik W = 0.95395 dengan p-value = 0.0002172. Karena p-value < 0.05, maka dapat disimpulkan bahwa residual tidak berdistribusi normal secara signifikan pada taraf kepercayaan 95%.

2. Uji Kolmogorov-Smirnov

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  rstandard(model_1y)
D = 0.12182, p-value = 0.04097
alternative hypothesis: two-sided
```

Gambar 30: Output dari Uji Shapiro-Wilk

Hasil uji Kolmogorov-Smirnov menunjukkan nilai D = 0.12182 dengan p-value = 0.04097. Sama seperti sebelumnya, p-value < 0.05, sehingga terdapat cukup bukti untuk menolak hipotesis nol sehingga residual tidak mengikuti distribusi normal secara signifikan.

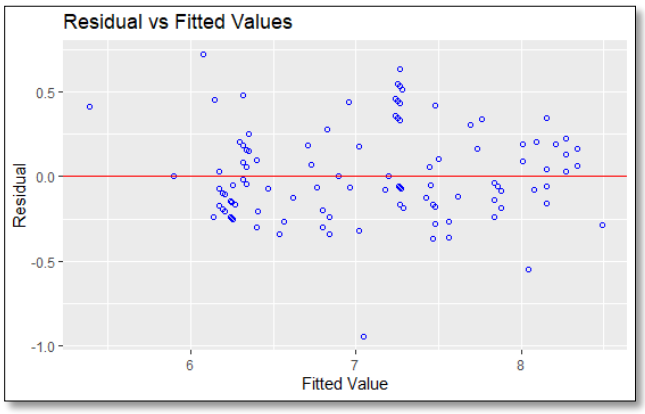
2. Uji Homoskedastisitas

```
# Uji Homoskedastisitas
ols_plot_resid_fit(model_akhir)
plot(fitted(model_akhir))
bptest(model_akhir)
```

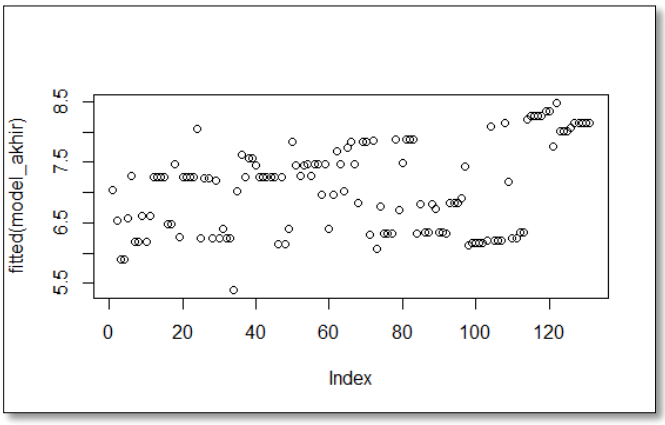
Gambar 31: Code Uji Homoskedastisitas

Uji homoskedastisitas bertujuan untuk mengevaluasi apakah varians dari residual dalam model regresi bersifat tetap (homoskedastik) atau mengalami perubahan (heteroskedastik). Asumsi ini merupakan bagian penting dalam model regresi linear untuk memastikan validitas hasil analisis. Pemeriksaan dilakukan melalui dua pendekatan, yaitu visual dan statistik.

- Pendekatan Visual



Gambar 32: Ouput berupa Plot Residual vs Fitted Values



Gambar 34: Output berupa Plot Fitted Model

Plot antara *residual* dan *fitted values* menunjukkan bahwa sebaran titik tidak mengikuti pola tertentu yang jelas. Walaupun terdapat sedikit ketidakteraturan dalam penyebarannya di beberapa bagian, secara keseluruhan titik-titik menyebar secara acak di sekitar garis horizontal nol. Hal ini memberikan indikasi awal bahwa gejala heteroskedastisitas tidak terlalu menonjol.

- Uji Statistik: Breusch-Pagan

```
> bptest(model_akhir)

        studentized Breusch-Pagan test

data:  model_akhir
BP = 43.64, df = 11, p-value = 8.4e-06
```

Gambar 35: Output dari Uji Statistik: Breusch-Pagan

Sebagai pelengkap dari analisis visual, dilakukan pengujian dengan metode Breusch-Pagan. Hasilnya menunjukkan nilai statistik uji sebesar BP

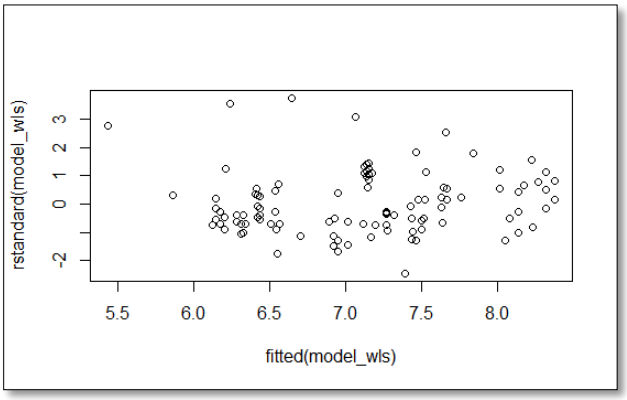
= 43.64 dengan p-value = 8.4e-06. Karena p-value lebih kecil dari 0.05, maka hipotesis nol yang menyatakan bahwa varians residual bersifat konstan ditolak pada tingkat signifikansi 5%. Oleh karena itu, terdapat dugaan kuat bahwa model mengalami masalah heteroskedastisitas.

Untuk mengatasi masalah heteroskedastisitas, dilakukan estimasi model ulang menggunakan metode Weighted Least Squares (WLS) sebagai berikut:

```
#WLS untuk Uji Homoskedastisitas
e3 <- e2^2
log_e3 <- log(e3)

model_var <- lm(log_e3 ~ I(Age^2) + I(Physical.Activity.Level^2) +
                  I(Stress.Level^2) + I(Heart.Rate^2) + Gender + BMI.Category +
                  Age:Heart.Rate + Quality.of.Sleep:Physical.Activity.Level +
                  Heart.Rate:Quality.of.Sleep, data = data)

pred_var <- exp(fitted(model_var))
data$w <- 1 / pred_var

model_wls <- lm(Sleep.Duration ~ I(Age^2) + I(Physical.Activity.Level^2) +
                  I(Stress.Level^2) + I(Heart.Rate^2) + Gender + BMI.Category +
                  Age:Heart.Rate + Quality.of.Sleep:Physical.Activity.Level +
                  Heart.Rate:Quality.of.Sleep, data = data, weights = w)


#Plot residual WLS
plot(fitted(model_wls), rstandard(model_wls))
bptest(model_wls)
```

Gambar 36: Code dari WLS untuk Uji Homoskedastisitas

- Visualisasi Hasil



Gambar 37: Ouput berupa Plot fitted(model_wls)

Gambar berikut menunjukkan plot antara nilai fitted values terhadap standardized residuals dari model regresi yang telah dikoreksi menggunakan metode Weighted Least Squares (WLS). Dengan demikian, plot ini memberikan bukti visual bahwa model WLS berhasil mengatasi masalah heteroskedastisitas yang sebelumnya ditemukan pada model awal (melalui uji Breusch-Pagan).

- Uji Statistik

```
> bptest(model_wls)

        studentized Breusch-Pagan test

data:  model_wls
BP = 678.49, df = 11, p-value < 2.2e-16
```

Gambar 38: Output dari *bptest(model_wls)*

Hasil uji menunjukkan nilai statistik BP = 678.49 dengan p-value < 2.2e-16. Karena p-value < 0.05, maka dapat disimpulkan bahwa varians residual bersifat konstan ditolak pada tingkat signifikansi 5%

3. Uji Independensi Residual

```
# Uji Independent
durbinWatsonTest(model_akhir)
```

Gambar 39: Code Uji Independent

Uji independensi dilakukan untuk mengetahui apakah sisa (residual) dari model regresi saling bebas atau justru saling berhubungan (autokorelasi). Salah satu cara yang umum digunakan untuk menguji hal ini adalah uji Durbin-Watson.

```
> durbinWatsonTest(model_akhir)
 lag Autocorrelation D-W Statistic p-value
   1       0.08367972      1.732392    0.05
 Alternative hypothesis: rho != 0
```

Gambar 40: Ouput dari Uji Independent

Berdasarkan hasil tersebut, dapat disimpulkan bahwa terdapat indikasi lemah terhadap autokorelasi positif pada residual model. Namun, karena p-value berada tepat di ambang 0.05, maka hasil ini bersifat marginal dan tidak memberikan bukti yang cukup kuat untuk menyatakan adanya autokorelasi secara signifikan.

4. Validasi Silang

```
# validasi silang (cross-validation)
train_control <- trainControl(method = "cv", number = 5)
model_akhir2 <- train(
  Sleep.Duration ~ I(Age^2) + I(Physical.Activity.Level^2) +
    I(Stress.Level^2) + I(Heart.Rate^2) + Gender + BMI.Category +
    Age:Heart.Rate + Quality.of.Sleep:Physical.Activity.Level +
    Heart.Rate:Quality.of.Sleep,
  data = data,
  method = "lm",
  trControl = train_control
)

model_akhir2
```

Gambar 41: Code Validisi Silang

Validasi silang (cross-validation) digunakan untuk mengevaluasi seberapa baik model regresi dapat memprediksi data. Teknik ini bekerja dengan membagi data menjadi beberapa bagian (disebut *fold*), lalu secara bergiliran model dilatih pada sebagian data dan diuji pada bagian lainnya. Dalam analisis ini digunakan 5-fold cross-validation, yang berarti data dibagi menjadi lima bagian dan proses pelatihan serta pengujian dilakukan sebanyak lima kali.

Model yang diuji terdiri dari 9 variabel prediktor dan telah melalui proses pra-pemrosesan (preprocessing). Hasil pengujian model menunjukkan:

```
> model_akhir2
Linear Regression

131 samples
  7 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 105, 105, 105, 105, 104
Resampling results:

  RMSE       Rsquared   MAE
  0.2741938  0.8756416  0.2230452

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Gambar 42: Output dari Validasi Silang

Nilai RMSE mencerminkan rata-rata besar kesalahan antara hasil prediksi model dengan data sebenarnya. Semakin kecil nilai RMSE, semakin akurat model. Dalam kasus ini, nilai RMSE sebesar 0.2741938 termasuk kecil, yang menunjukkan bahwa model memiliki performa prediksi yang cukup baik.

Sementara itu, nilai R^2 sebesar 0.8756416 berarti bahwa sekitar 87,56% variasi dari variabel target (*Sleep Duration*) bisa dijelaskan oleh variabel-variabel bebas dalam model. Ini menunjukkan bahwa model memiliki hubungan yang kuat terhadap data.

Untuk mendapatkan nilai Mean Squared Error (MSE) adalah sebagai berikut:

```
# Melihat nilai RMSE dari hasil validasi silang (cross-validation)
rmse_val <- model_akhir2$results$RMSE
mse_val <- rmse_val^2
mse_val
```

Gambar 43: Code Mengecek Nilai MSE

```
> mse_val
[1] 0.07518226
```

Gambar 44: Ouput untuk Nilai MSE

Nilai MSE sebesar 0.07518226 menunjukkan bahwa model menghasilkan tingkat kesalahan prediksi yang rendah, sehingga model yang dibangun dapat dianggap cukup andal dan layak digunakan untuk keperluan analisis atau prediksi ke depan.

BAB V

PENUTUP

5.1 Kesimpulan

Berdasarkan hasil evaluasi melalui analisis visual dan pengujian statistik, dapat diketahui bahwa model regresi linier ini secara umum memiliki kemampuan prediksi yang tergolong baik, walaupun masih terdapat beberapa pelanggaran asumsi klasik.

Dilihat dari aspek normalitas residual, hasil histogram dan Q-Q Plot yang telah ditransformasi pada variabel dependen menunjukkan bahwa distribusi residual belum sepenuhnya mengikuti pola normal, yang juga diperkuat dengan hasil uji Shapiro-Wilk dan Kolmogorov-Smirnov. Sementara itu, uji Breusch-Pagan menunjukkan adanya indikasi masalah heteroskedastisitas pada model, sehingga varians residual belum homogen sepenuhnya.

Selain itu, pengujian independensi residual menggunakan uji Durbin-Watson menghasilkan nilai statistic sebesar 1.732392 dengan p-value 0.05, yang berarti indikasi lemah terhadap autokorelasi positif pada residual model. Namun, karena p-value berada tepat di ambang 0.05, maka hasil ini bersifat marginal dan tidak memberikan bukti yang cukup kuat untuk menyatakan adanya autokorelasi secara signifikan.

Walaupun terdapat pelanggaran pada normalitas, homoskedastisitas, dan independensi residual, performa prediksi model tetap dinilai cukup baik. Hal ini tercermin dari nilai Adjusted R-squared sebesar 0.8756416, yang berarti model mampu menjelaskan sekitar 87,56% variasi durasi tidur melalui variabel-variabel prediktor yang digunakan. Nilai galat prediksi pun tergolong kecil, terlihat dari RMSE sebesar 0.2741938 dan MSE sebesar 0.07518226.

Secara keseluruhan, meskipun tidak seluruh asumsi regresi terpenuhi secara ideal, model regresi linier ini tetap dapat digunakan untuk keperluan prediksi durasi tidur dengan tingkat ketepatan yang relatif memadai.

5.2 Saran

Berdasarkan hasil analisis dan evaluasi model regresi linear yang telah dilakukan, terdapat beberapa saran yang dapat diberikan untuk pengembangan penelitian lebih lanjut maupun perbaikan teknis:

1. Peningkatan Pemenuhan Asumsi Klasik

   Hasil evaluasi menunjukkan adanya pelanggaran pada beberapa asumsi klasik regresi, terutama pada normalitas residual dan homoskedastisitas. Untuk itu, disarankan agar penelitian selanjutnya mengeksplorasi transformasi data yang lebih optimal (seperti Box-Cox atau logaritma), atau mempertimbangkan penggunaan metode regresi robust seperti Quantile Regression atau Generalized Least Squares yang lebih toleran terhadap pelanggaran asumsi.

2. Pengayaan Variabel Prediktor

   Penelitian ini berfokus pada variabel-variabel fisiologis dan gaya hidup seperti usia, detak jantung, stres, dan aktivitas fisik. Ke depan, penelitian serupa dapat mempertimbangkan faktor lain yang relevan secara psikologis maupun lingkungan, seperti tingkat konsumsi kafein, waktu paparan layar (screen time), atau kualitas lingkungan tidur, untuk memperkaya model prediksi durasi tidur.

3. Penggunaan Dataset yang Lebih Besar dan Variatif

   Dataset yang digunakan hanya mencakup 374 observasi. Dengan ukuran sampel yang lebih besar dan bervariasi secara demografis (misalnya berdasarkan wilayah, pekerjaan, atau rentang usia yang lebih luas), model yang dibangun akan memiliki kekuatan generalisasi yang lebih baik.

4. Evaluasi Model Tambahan

   Walaupun validasi silang 5-fold telah digunakan, disarankan untuk menggunakan metrik evaluasi tambahan seperti MAE (Mean Absolute Error) atau AUC (jika klasifikasi digunakan di masa mendatang). Hal ini dapat memberikan pandangan yang lebih komprehensif terhadap performa model.

5. Visualisasi Interaktif dan Interpretasi Model

   Untuk meningkatkan interpretabilitas hasil regresi terutama bagi non-statistisi, disarankan untuk menyertakan visualisasi interaktif seperti effect plots, partial

dependence plots, atau menggunakan Shiny App di R untuk menampilkan hasil interaktif dari pengaruh variabel terhadap prediksi durasi tidur.

6. Integrasi dengan Pendekatan Machine Learning

   Di masa depan, pendekatan regresi linear dapat dibandingkan dengan metode machine learning seperti Random Forest atau Gradient Boosting untuk melihat apakah model yang lebih kompleks mampu memberikan peningkatan signifikan dalam akurasi prediksi tanpa mengorbankan interpretabilitas

REFERENSI

Mendenhall, W. dan Sincich, T. (2003). *A Second Course in Statistics: Regression Analysis*. Edisi ke-7. Boston: Pearson Education, Inc.

APPENDIKS

https://colab.research.google.com/drive/14ndgX5OTP_jPDH_OAxe0oP3GUInSYivu?usp=sharing