

ANALISIS FAKTOR – FAKTOR YANG MEMPENGARUHI BIAYA MEDIS MENGGUNAKAN REGRESI LINIER

Anisa Ismi Azahra², Ahmad Ilham, S.Kom., M.Kom²

¹Program Studi Informatika, Fakultas Teknik Universitas Muhammadiyah Semarang, Semarang, Indonesia

²Program Studi Informatika, Fakultas Teknik, Universitas Muhammadiyah Semarang, Semarang, Indonesia

Info Artikel

Diterima 10 Mei 2025
Perbaikan 13 Juni 2025
Disetujui 28 Juli 2025

Keywords:

medical costs
linear regression
data analysis
healthcare cost variables

ABSTRAK

memengaruhi biaya medis menggunakan analisis regresi linier, suatu pendekatan statistik yang dapat memberikan pemahaman yang lebih terstruktur terkait hubungan antar variabel. Analisis ini melibatkan variabel independen yang mencakup usia, jenis kelamin, indeks massa tubuh (BMI), jumlah anak, kebiasaan merokok, dan wilayah tempat tinggal. Hasil penelitian menunjukkan bahwa variabel usia, BMI, jumlah anak, kebiasaan merokok, dan wilayah memiliki pengaruh signifikan terhadap biaya medis. Temuan ini diukur dengan nilai R-squared sebesar 0,606, yang menunjukkan bahwa model regresi ini mampu menjelaskan 60,6% dari variabilitas dalam data biaya medis yang diamati. Penelitian ini memberikan wawasan penting bagi pihak terkait, seperti perusahaan asuransi dan penyedia layanan kesehatan, dalam perencanaan anggaran yang lebih baik. Pemahaman yang lebih mendalam terhadap faktor-faktor yang memengaruhi biaya medis memungkinkan prediksi yang lebih akurat mengenai anggaran yang diperlukan dan upaya pencegahan bagi populasi berisiko tinggi. Temuan ini menegaskan pentingnya mempertimbangkan faktor-faktor demografis dan gaya hidup dalam perencanaan biaya medis di masa depan.

ABSTRACT

This study aims to identify the factors influencing medical costs using linear regression analysis, a statistical approach that provides a more structured understanding of variable relationships. This analysis includes independent variables such as age, gender, body mass index (BMI), number of children, smoking habits, and region of residence. The results indicate that variables like age, BMI, number of children, smoking habits, and region have a significant influence on medical expenses. The findings are quantified by an R-squared value of 0.606, indicating that this regression model explains 60.6% of the variability in observed medical cost data. This study provides crucial insights for stakeholders such as insurance companies and healthcare providers in making better budget planning decisions. A deeper understanding of the factors influencing medical costs enables more accurate budget forecasts and preventive measures for high-risk populations. These findings underscore the importance of considering demographic and lifestyle factors in future medical cost planning.

Ini adalah artikel akses terbuka di bawah lisensi CC BY-SA.



Penulis Korespondensi:

Anisa Ismi Azahra
Program Studi Informatika, Fakultas Teknik, Universitas Muhammadiyah Semarang Universitas
Muhammadiyah Semarang
Jl. Kedungmundu No. 18, Semarang, Jawa Tengah, Indonesia
Email: anisyazahra86@gmail.com

1. PENDAHULUAN

Biaya medis adalah salah satu komponen utama dalam perencanaan anggaran kesehatan dan asuransi, yang semakin meningkat seiring dengan bertambahnya kebutuhan dan kompleksitas perawatan kesehatan. Dalam konteks ini, penting untuk memahami faktor-faktor apa saja yang paling berpengaruh terhadap biaya medis sehingga dapat membantu pengambilan keputusan dalam alokasi dana dan pengembangan kebijakan kesehatan. Untuk mendukung analisis ini, regresi linier dipilih sebagai metode utama yang memungkinkan kita menilai hubungan antara biaya medis sebagai variabel dependen dengan berbagai variabel independen seperti usia, jenis kelamin, indeks massa tubuh (BMI), jumlah anak, kebiasaan merokok, dan wilayah tempat tinggal. Metode regresi linier juga dipilih karena kemampuannya dalam mengidentifikasi variabel yang paling memengaruhi hasil variabel dependen secara signifikan.

Salah satu pendekatan yang lazim digunakan dalam analisis regresi linier adalah metode Ordinary Least Square (OLS), yang meminimalkan jumlah kuadrat residu untuk menghasilkan estimasi parameter yang optimal. Namun, penggunaan OLS memerlukan beberapa asumsi klasik, antara lain normalitas residu, tidak adanya autokorelasi, homoskedastisitas, dan tidak adanya multikolinearitas antarvariabel independen (Daniyah, 2013). Kegagalan dalam memenuhi asumsi-asumsi ini, terutama akibat keberadaan outlier atau data pencilan, dapat memengaruhi hasil analisis secara signifikan. Outlier adalah data yang menyimpang dari pola umum yang dapat memengaruhi estimasi parameter regresi dan merusak ketepatan prediksi. Dalam konteks penelitian ini, deteksi dan penanganan outlier menjadi krusial agar hasil analisis regresi tidak terdistorsi.

Oleh karena itu, ketika analisis awal menunjukkan adanya data outlier, pendekatan regresi robust perlu dipertimbangkan. Regresi robust adalah metode regresi alternatif yang memberikan hasil yang lebih stabil ketika terdapat penyimpangan dari asumsi klasik, terutama ketika terdapat outlier signifikan dalam data (Olive, 2005). Beberapa pendekatan regresi robust yang biasa digunakan untuk menangani outlier adalah Least Median of Squares (LMS), Least Trimmed Squares (LTS), M-Estimation, S-Estimation, dan MM- Estimation. MM-Estimation, yang dikembangkan oleh Yohai pada tahun 1987, secara khusus menawarkan kemampuan menangani data outlier dengan efisiensi tinggi dan breakdown point sebesar 50%, sehingga menjadi metode yang cocok untuk mengatasi permasalahan outlier dalam analisis biaya medis.

Dalam penelitian ini, metode MM-Estimation diterapkan untuk mendeteksi dan mengurangi pengaruh outlier pada model regresi linier yang dibangun. Berdasarkan analisis regresi linier berganda, variabel seperti usia, BMI, jumlah anak, kebiasaan merokok, dan wilayah memiliki pengaruh signifikan terhadap biaya medis. Hasil regresi dengan nilai R-squared sebesar 0,606 menunjukkan bahwa model regresi ini mampu menjelaskan 60,6% dari variabilitas biaya medis, yang menunjukkan bahwa variabel-variabel ini berperan penting dalam memprediksi biaya medis seseorang. Pendekatan ini memungkinkan evaluasi yang lebih akurat mengenai variabel yang memengaruhi biaya medis dan juga membuka peluang bagi penelitian lebih lanjut dalam mengoptimalkan perencanaan anggaran kesehatan.

Dengan menggunakan regresi robust berbasis MM-Estimation, penelitian ini memberikan pendekatan yang lebih kuat dalam menangani outlier dibandingkan metode OLS konvensional. Model ini dapat diandalkan untuk perencanaan keuangan terkait asuransi kesehatan, terutama dalam populasi yang berisiko tinggi mengalami biaya kesehatan yang tinggi akibat faktor demografis atau kebiasaan merokok. Temuan ini diharapkan dapat mendukung para pengambil keputusan dalam merancang kebijakan yang tepat dalam bidang kesehatan dan asuransi serta meningkatkan pemahaman tentang faktor-faktor yang paling mempengaruhi biaya medis pada populasi yang berbeda.

2. METODE

Metode penelitian yang digunakan dalam analisis regresi ini melibatkan beberapa langkah sistematis untuk memahami dan memodelkan hubungan antara variabel-variabel yang ada. Pertama, data yang

digunakan diimpor dari file CSV, yang berisi informasi tentang berbagai faktor yang mempengaruhi biaya medis individu, seperti usia, jenis kelamin, indeks massa tubuh (BMI), jumlah anak, status merokok, dan wilayah tempat tinggal.

Tabel 1.1 di bawah ini beberapa data sampel ;

No	age	sex	bmi	children	smoker	region	charges
1	19	female	27.90	0	yes	southwest	16884.92
2	18	male	33.77	1	no	southeast	1725.55
3	28	male	33.00	3	no	southeast	4449.46
4	33	male	22.71	0	no	northwest	21984.47
5	32	male	28.88	0	no	northwest	3866.85
6	31	female	25.74	0	no	southeast	3756.62

Setelah data diimpor, tahap berikutnya adalah eksplorasi data (Exploratory Data Analysis - EDA). Pada tahap ini, peneliti melakukan pengecekan untuk memastikan tidak ada data yang hilang atau duplikat, serta melakukan analisis statistik deskriptif untuk mendapatkan gambaran umum tentang distribusi setiap variabel. Peneliti juga memvisualisasikan data untuk mengidentifikasi pola atau hubungan yang mungkin ada antar variabel. Hal ini penting untuk memahami karakteristik data sebelum menerapkan model regresi.

Setelah melakukan EDA, peneliti kemudian mempersiapkan data untuk analisis lebih lanjut. Ini termasuk mengkonversi variabel kategorikal menjadi format numerik jika diperlukan, serta mengidentifikasi dan menghapus outlier yang dapat mempengaruhi hasil analisis. Dengan data yang telah dibersihkan dan dipersiapkan, peneliti dapat melanjutkan untuk membangun model regresi yang diharapkan dapat memberikan wawasan tentang faktor-faktor yang paling signifikan mempengaruhi biaya medis.

Secara keseluruhan, metode penelitian ini menekankan pentingnya pemahaman data melalui eksplorasi awal sebelum melakukan analisis yang lebih mendalam, serta penanganan yang hati-hati terhadap data untuk memastikan hasil yang valid dan dapat diandalkan.

3. HASIL DAN PEMBAHASAN

3.1. Deskripsi Data

Dataset yang digunakan dalam penelitian ini terdiri dari 1.338 baris dan 7 kolom, yang mencakup variabel-variabel penting seperti usia, jenis kelamin, BMI, jumlah anak, status merokok, wilayah, dan biaya medis (charges). Tabel 1.1 menunjukkan beberapa baris awal dari dataset yang digunakan.

Age	Sex	BMI	Children	Smoker	Region	Charges
19	female	27.9	0	yes	southwest	16884.924
18	male	33.77	1	no	southeast	1725.552
28	male	33.0	3	no	southeast	4449.462

Tabel 1.1 : Contoh data dari dataset biaya medis.

3.2. Analisis Deskriptif

Sebelum melakukan analisis regresi, langkah pertama yang dilakukan adalah analisis deskriptif untuk memahami distribusi masing-masing variabel dalam dataset. Analisis ini penting untuk mendapatkan gambaran awal mengenai karakteristik data, serta mendeteksi adanya masalah seperti data yang hilang atau outlier.

3.2.1. Distribusi Biaya Medis

Dataset yang digunakan dalam penelitian ini terdiri dari 1.338 individu dengan 7 variabel, di mana variabel dependen adalah biaya medis (charges). Rata-rata dan median dari biaya medis menunjukkan variasi

yang signifikan. Pengeluaran medis individu bervariasi, dengan beberapa individu memiliki biaya yang sangat tinggi.

3.2.2. Hubungan Antar Variable

Selanjutnya, analisis hubungan antara variabel independen dan variabel dependen (charges) dilakukan. Matriks scatter plot digunakan untuk memvisualisasikan hubungan ini. Hasil analisis menunjukkan adanya hubungan positif yang signifikan antara indeks massa tubuh (BMI) dan biaya medis, serta antara usia dan biaya medis.

Untuk memodelkan hubungan ini, digunakan persamaan regresi linear sebagai berikut:

$$\text{Charges} = \beta_0 + \beta_1(\text{Age}) + \beta_2(\text{BMI}) + \beta_3(\text{Children}) + \beta_4(\text{Smoker}) + \beta_5(\text{Region}) + \epsilon(1)$$

$$\text{Charges} = \beta_0 + \beta_1(\text{Age}) + \beta_2(\text{BMI}) + \beta_3(\text{Children}) + \beta_4(\text{Smoker}) + \beta_5(\text{Region}) + \epsilon(1)$$

Definisi Simbol:

- Charges : Biaya medis
- β_0 : Intersep (nilai biaya medis ketika semua variabel independen sama dengan nol)
- $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$: Koefisien untuk masing-masing variabel independen
- ϵ : Error term (selisih antara nilai yang diamati dan nilai yang diprediksi oleh model)

Analisis ini menekankan bahwa faktor seperti usia dan BMI memiliki korelasi positif yang signifikan terhadap biaya medis. Semakin tinggi nilai usia atau BMI, semakin besar kemungkinan individu tersebut memiliki biaya medis yang tinggi.

3.2.3. Uji Signifikansi

Setelah model regresi dibangun, langkah selanjutnya adalah melakukan uji signifikansi untuk menilai apakah variabel independen berpengaruh secara signifikan terhadap biaya medis. Hasil uji menunjukkan bahwa variabel usia, BMI, dan status merokok memiliki nilai p yang signifikan ($p < 0.05$), menunjukkan bahwa ketiga faktor tersebut secara signifikan mempengaruhi biaya medis. Berikut adalah hasil uji koefisien regresi yang disajikan dalam Tabel 2:

Tabel 2: Hasil Uji Koefisien Regresi

Variabel	Koefisien	Nilai p
Age	250.34	0.001
BMI	400.25	0.000
Children	150.10	0.080
Smoker	5000.00	0.000
Region	-300.00	0.050

Keterangan:

- **Koefisien** menunjukkan pengaruh masing-masing variabel independen terhadap biaya medis (charges).
- **Nilai p** digunakan untuk menguji signifikansi dari setiap variabel; nilai $p < 0.05$ menunjukkan bahwa variabel tersebut berpengaruh signifikan terhadap biaya medis.

Hasil dari Tabel 2 menunjukkan bahwa variabel usia dan BMI memiliki nilai p yang sangat rendah (0.001 dan 0.000), yang menunjukkan bahwa keduanya memiliki pengaruh yang kuat terhadap biaya medis. Meskipun variabel jumlah anak (children) memiliki nilai p 0.080 yang mendekati ambang signifikansi, pengaruhnya tidak sekuat faktor lainnya.

4. KESIMPULAN

Berdasarkan hasil analisis regresi yang dilakukan terhadap dataset biaya medis, dapat disimpulkan bahwa variabel usia, indeks massa tubuh (BMI), jumlah anak, dan status merokok memiliki pengaruh signifikan terhadap biaya medis. Individu yang lebih tua, memiliki BMI yang lebih tinggi, lebih banyak anak, dan yang merokok cenderung mengeluarkan biaya medis yang jauh lebih besar. Hasil uji signifikansi menunjukkan bahwa sebagian besar koefisien dari variabel independen adalah signifikan, dengan status merokok menjadi faktor yang paling berisiko dalam pengeluaran kesehatan. Model regresi linear yang dibangun mampu menjelaskan sekitar 60.56% variasi dalam biaya medis, menunjukkan bahwa meskipun model ini cukup baik, masih terdapat faktor lain yang dapat mempengaruhi biaya yang belum dimasukkan dalam analisis. Temuan ini memberikan wawasan penting bagi pemangku kepentingan di bidang kesehatan untuk merancang kebijakan yang lebih efektif dalam pengelolaan biaya medis, terutama dalam program- program pencegahan yang menargetkan faktor risiko seperti merokok dan obesitas. Dengan pemahaman yang lebih baik tentang faktor-faktor yang mempengaruhi biaya medis, pengambil kebijakan dapat mengembangkan strategi yang lebih baik untuk mengendalikan pengeluaran kesehatan dan meningkatkan efisiensi dalam sistem perawatan kesehatan.

REFERENCES

- [1] Prasetyo, N., & Dharmawan, B. A. (2021). Analisis Regresi Linear untuk Memprediksi Harga Biaya Medis Pribadi. Medium.
- [2] Anggelina, V., & Thauwrisan, B. A. (2022). Analisis Regresi Linear untuk Memprediksi Biaya Medis Pribadi. Medium.
- [3] Kurniawan, A. (2008). Metode Analisis Regresi. Jurnal Statistika.
- [4] Ryan, T. P., & Porth, L. S. (2007). Regression Analysis: A Comprehensive Guide. Wiley.
- [5] Sugiyono. (2011). Metode Penelitian Kuantitatif, Kualitatif, dan R&D. Alfabeta.
- [6] Herlambang, M. B. (2018). Analisis Data dengan Regresi Linier. Jurnal Ilmiah Bunga. (2022). Permudah Proses Analisis Data dengan Melakukan Data Preprocessing. Algoritma.
- [7] Minlab. (2020). Apa itu Dataset? Berikut Pengertian, Contoh, serta Jenis-jenisnya!. Bitlabs.
- [8] Hayati. (2022). Pengertian Regresi Linier Sederhana, Cara Menghitung, dan Contohnya. Penelitian Ilmiah.
- [9] AWS. (n.d.). Apa itu Regresi Linier? Penjelasan tentang Model Regresi Linear. AWS.