BAB I
PENDAHULUAN

1.1 Latar Belakang

Regresi linier merupakan metode statistik yang digunakan untuk membangun model hubungan antara variabel dependen dengan variabel independen. Regresi linier terbagi menjadi dua bagian, yaitu regresi linier sederhana dan regresi linier berganda. Model regresi linier berganda digunakan untuk menganalisis hubungan antara satu variabel dependen dengan lebih dari satu variabel independen. Pendugaan parameter dalam regresi linier umumnya dilakukan menggunakan metode *Ordinary Least Squares* (OLS), yang bertujuan untuk meminimalkan jumlah kuadrat selisih antara nilai aktual dan nilai prediksi. Sebelum model regresi digunakan, terdapat beberapa asumsi yang harus dipenuhi, yaitu kenormalan residual, tidak adanya multikolinieritas, homoskedastisitas, dan tidak terjadi autokorelasi. Salah satu kemungkinan penyebab tidak terpenuhinya asumsi-asumsi tersebut adalah karena adanya data *outlier*, yang dapat memberikan dampak signifikan terhadap hasil estimasi parameter dan analisis regresi, terutama jika pengamatan tersebut memiliki pengaruh yang signifikan terhadap model [1].

Outlier merupakan pengamatan yang menyimpang dari pola data utama. Dalam analisis regresi, *outlier* diidentifikasi sebagai pengamatan yang tidak sesuai dengan model yang digunakan, yang biasanya ditandai oleh nilai residual yang besar, yaitu perbedaan signifikan antara nilai aktual yang diamati dengan nilai prediksi yang dihasilkan oleh model. Keberadaan *outlier* mengindikasikan bahwa pengamatan tersebut tidak sepenuhnya dijelaskan oleh model. *Outlier* dapat ditemukan dalam data nyata dan sering kali tidak terdeteksi karena proses pengolahan data yang dilakukan secara otomatis tanpa pemeriksaan mendalam. Keberadaan *outlier* dapat disebabkan oleh berbagai faktor, seperti kesalahan input data, misalnya kesalahan pengetikan, penempatan titik desimal, kesalahan pencatatan, transmisi data, serta masuknya anggota populasi yang berbeda secara tidak sengaja ke dalam sampel [2]-[3].

Penelitian yang dilakukan oleh S. Indra [4], menyatakan bahwa keberadaan *outlier* dapat menyebabkan suatu pengamatan dikategorikan sebagai *high leverage point* atau *influential observation*, yang pada akhirnya dapat mengganggu validitas model regresi. Melalui penerapan metode diagnostik, seperti *Studentized Residual* dan *DFFITS*, studi ini mengungkap bahwa keberadaan *outlier* dapat berdampak signifikan terhadap hasil analisis regresi. Dampak tersebut mencakup perubahan nilai *intercept*, *slope*, koefisien determinasi (R^2), serta varians *error* (S^2), yang secara keseluruhan dapat memengaruhi interpretasi dan kesimpulan yang diperoleh dari model regresi.

Dalam analisis regresi linier berganda, keberadaan *multiple outliers* (terdapat lebih dari satu *outlier* dalam dataset) dapat memberikan dampak serius terhadap hasil analisis. jika *multiple outliers* tidak dideteksi dan ditangani dengan baik, maka model regresi yang dihasilkan tidak akan mampu memberikan gambaran yang akurat mengenai hubungan antara variabel-variabel yang dianalisis. Oleh karena itu, pendeteksian *multiple outliers* menjadi langkah penting untuk memastikan keakuratan serta validitas model regresi yang digunakan.

Faktanya, fenomena keberadaan *outlier* ini sering ditemukan dalam analisis data empiris. Misalnya dalam penelitian sebelumnya yang membahas pendeteksian *multiple outliers* menggunakan analisis residual dan perhitungan subset dalam metode *Internally Studentized Residual*, ditemukan beberapa pengamatan *outlier* dalam dataset *stack-loss* (sebuah penelitian yang meneliti faktor-faktor yang memengaruhi proses penyerapan nitrogen oksida selama 21 hari operasi pabrik) [2].

Penelitian mengenai deteksi *multiple outlier* dalam regresi linier berganda telah dikaji dalam berbagai konteks dan metode. Sebagai contoh, studi yang dilakukan oleh U. T. Ejiofor *et al*. [2], mengembangkan metode pendeteksian *multiple outliers* dengan pendekatan statistik uji berbasis *subset,* yang menerapkan metode *Internally Studentized Residuals*. Pendekatan ini menawarkan solusi alternatif yang lebih efisien dibandingkan dengan metode pendeteksian pada umumnya yang memerlukan penghapusan pengamatan atau penyusunan ulang model pada setiap tahap analisis, seperti yang dilakukan pada penelitian [5] mengenai metode pendeteksian *multiple outliers*, prosedur identifikasi dilakukan

secara bertahap dengan mengeliminasi pengamatan yang paling ekstrem berdasarkan nilai *Studentized Residual* terbesar. Proses ini kemudian diulang pada data yang tersisa hingga hipotesis yang menyatakan tidak adanya *outlier* diterima.

Selain itu, penelitian yang dilakukan oleh E. I. Mba *et al.* [6], membahas identifikasi data *outlier* dalam bentuk *subset* pada model regresi linier. Perbedaannya terletak pada pendekatan yang digunakan, yaitu metode *Bounded Index Plot*, yang memperkenalkan pendekatan berbasis matriks diagnostik *T*. Pendekatan tersebut memungkinkan identifikasi *subset* yang mengandung *outlier* dengan menganalisis elemen-elemen di luar diagonal utama dalam matriks tersebut (elemen off-diagonal).

Dalam konteks analisis sosial dan ekonomi, regresi linier berganda sering dimanfaatkan untuk meneliti faktor-faktor yang memengaruhi tingkat kemiskinan di suatu wilayah. Misalnya, pada penelitian yang dilakukan oleh R. Aristiarto *et al*. [7], menunjukkan bahwa data indeks keparahan kemiskinan di Indonesia pada tahun 2021 mengandung *outlier*, yang menyebabkan asumsi normalitas tidak terpenuhi. Akibatnya, metode regresi OLS (*Ordinary Least Squares*) menjadi kurang optimal dalam analisis ini. Untuk mengatasi kendala tersebut, penelitian ini menerapkan estimasi *Generalized M (GM)*, yang bertujuan meningkatkan ketahanan model terhadap pengaruh *outlier*. Hasil penelitian menunjukkan bahwa beberapa faktor yang secara signifikan memengaruhi indeks keparahan kemiskinan adalah persentase penduduk miskin, Indeks Pembangunan Manusia (IPM), serta proporsi rumah tangga dengan status kepemilikan rumah sendiri.

Kemiskinan merupakan suatu kondisi di mana individu atau kelompok masyarakat mengalami keterbatasan dalam aspek ekonomi, sehingga tidak mampu memenuhi kebutuhan dasar mereka. Isu kemiskinan merupakan permasalahan yang kompleks karena dapat dipengaruhi oleh berbagai faktor, seperti meningkatnya angka pengangguran, ketimpangan dalam Indeks Pembangunan Manusia (IPM), serta rendahnya pengeluaran per kapita untuk kebutuhan pangan. Selain itu, kemiskinan juga dapat disebabkan oleh pertumbuhan penduduk yang tidak diimbangi dengan perkembangan ekonomi yang memadai, yang pada akhirnya berdampak pada menurunnya kesejahteraan masyarakat [8].

Dalam penelitian ini, model regresi digunakan untuk menganalisis hubungan antara faktor-faktor yang memengaruhi kesmiskinan dengan tingkat kemiskinan di Provinsi Jawa Barat tahun 2024. Namun, karena data ekonomi dan sosial sering kali memiliki variabilitas yang tinggi serta berpotensi mengandung pengamatan ekstrem, kemungkinan adanya *outlier* dalam dataset cukup besar. Oleh karena itu, deteksi *multiple outliers* menjadi aspek krusial dalam memastikan bahwa model regresi yang dibangun mampu menghasilkan estimasi yang valid dan dapat diandalkan. Dengan demikian, model regresi yang telah bebas dari pengaruh *outlier* diharapkan dapat menghasilkan estimasi yang lebih akurat, sehingga hasil analisisnya dapat dijadikan landasan yang andal dalam perumusan kebijakan yang tepat sasaran dan berbasis data.

Penelitian ini akan berfokus pada pendeteksian *multiple outliers* dalam model regresi linier berganda menggunakan metode *Studentized Residual*, yaitu *Internally Studentized Residual* dan *Externally Studentized Residual*. Metode tersebut memungkinkan untuk mengidentifikasi pengamatan yang memiliki pengaruh signifikan terhadap model regresi. Dengan menerapkan metode tersebut pada data kemiskinan di Provinsi Jawa Barat tahun 2024, penelitian ini diharapkan dapat memberikan pemahaman yang lebih mendalam mengenai keberadaan *outlier* dalam analisis regresi.

1.2 Rumusan Masalah

Keberadaan *multiple outliers* dalam analisis regresi linier berganda dapat memengaruhi validitas hasil estimasi secara signifikan. Oleh karena itu, diperlukan langkah-langkah metode yang dapat digunakan untuk mengidentifikasi data yang termasuk sebagai *outlier* untuk meningkatkan keakuratan hasil analisis regresi. Metode *Studentized Residual* digunakan untuk mendeteksi *multiple outliers* dengan mengukur nilai residual terbesar yang diduga sebagai *outlier* dalam data kemiskinan di Provinsi Jawa Barat tahun 2024.

1.3 Batasan Masalah

Permasalahan dalam penelitian ini dibatasi oleh beberapa hal, di antaranya adalah sebagai berikut:

1. Penelitian ini hanya akan berfokus pada pendeteksian *multiple outliers* menggunakan dua metode, yaitu *Internally Studentized Residual* dan *Externally Studentized Residual* dalam konteks model regresi linier berganda
2. Data yang digunakan dalam penelitian ini adalah data kemiskinan di Provinsi Jawa Barat tahun 2024 yang diperoleh dari situs resmi Badan Pusat Statistik (BPS) Provinsi Jawa barat
3. Proses deteksi *outlier* akan dilakukan menggunakan bahasa pemrograman *Python* melalui *platform Google Colab*, serta menggunakan perangkat lunak SPSS untuk mendukung analisis statistik.

1.4 Tujuan Penelitian

Penelitian ini bertujuan untuk mendeteksi *multiple outliers* dalam analisis regresi linier berganda dengan metode *Studentized Residual* pada data kemiskinan di Provinsi Jawa Barat tahun 2024. Melalui penelitian ini, diharapkan dapat diketahui pengamatan yang menyimpang secara signifikan dari pola umum data yang dapat memengaruhi keakuratan hasil analisis data.

1.5 Metode Penelitian

Metode yang digunakan dalam penelitian ini adalah sebagai berikut:

1. Studi Literatur

   Tahap ini bertujuan untuk mengumpulkan berbagai teori, data, dan informasi yang berkaitan dengan regresi linier berganda, estimasi parameter, *outlier*, dan metode identifikasi *outlier.* Sumber referensi yang digunakan meliputi buku, jurnal ilmiah, artikel, serta literatur relevan lainnya.
2. Analisis dan Simulasi

   Pada tahap ini, peneliti mengevaluasi dan menganalisis hasil yang diperoleh dari tahap studi literatur, menyesuaikannya dengan permasalahan yang dikaji dalam penelitian ini. Selanjutnya, peneliti melakukan simulasi untuk proses identifikasi *multiple outliers* dalam model regresi linier berganda. Simulasi

ini diterapkan pada data kemiskinan di Provinsi Jawa Barat tahun 2024 dengan memanfaatkan bahasa pemrograman *Python* melalui *platform Google Colab*, serta menggunakan perangkat lunak SPSS.

3. Kesimpulan

   Pada tahap ini, peneliti merumuskan kesimpulan berdasarkan hasil analisis yang diperoleh dari proses simulasi yang telah dilakukan.

1.6 Sistematika Penulisan

Penelitian ini disusun berdasarkan sistematika penulisan yang terdiri dari lima bab utama, dengan rincian sebagai berikut:

BAB I PENDAHULUAN

Bab ini menyajikan pendahuluan dari penelitian yang mencakup latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, metode penelitian, dan sistematika penulisan.

BAB II LANDASAN TEORI

Bab ini memaparkan teori-teori yang relevan dengan topik penelitian, meliputi analisis regresi, asumsi regresi klasik, analisis residual, metode kuadrat terkecil, pencilan (*outlier*), *multiple outliers*, identifikasi *outlier*, serta pendekatan *bonferroni*. Teori-teori ini bertujuan untuk memberikan pemahaman menyeluruh terkait dasar-dasar teoritis yang mendukung penelitian ini.

BAB III PENDETEKSIAN *MULTIPLE OUTLIERS* DALAM MODEL REGRESI LINIER BERGANDA

Bab ini menjelaskan prosedur pendeteksian *multiple outliers* menggunakan metode *Internally Studentized Residual* dan *Externally Studentized Residual* pada data regresi linier berganda, penjelasan tersebut mencakup langkah-langkah sistematis pendeteksian *multiple outliers* dan rumus-rumus yang akan digunakan dalam perhitungan metode yang digunakan.

BAB IV STUDI KASUS DAN ANALISA

Bab ini berisi studi berdasarkan penelitian yang dilakukan, kemudian diaplikasikan dalam data kemiskinan di Provinsi Jawa Barat tahun 2024, dengan interpretasi dari hasil analisis yang diperoleh.

BAB V PENUTUP

Bab ini berisi kesimpulan dari hasil penelitian yang telah dilakukan, serta saran yang dapat dikembangkan lebih lanjut untuk memperbaiki dan memperluas cakupan penelitian di masa mendatang.