BAB I

PENDAHULUAN

1.1 Latar Belakang

Analisis regresi merupakan metode yang sering digunakan pada suatu penelitian dalam bidang statistika. Analisis regresi mempunyai tahapan yang harus dikerjakan salah satunya mengestimasi parameter dengan menggunakan *Ordinary Least Square* (OLS) yaitu proses penaksiran parameter regresi dengan cara mencari nilai minimum dari penjumlahan kuadrat kesalahan (*error*) pada bentuk persamaan regresi [1].

Pada umumnya, pasti ada beberapa data yang memang melenceng daripada perkiraan yang sebelumnya telah ditetapkan. Untuk mencari atau menganalisis data yang kemungkinan melenceng atau keluar dari pola umum model yang sudah ditetapkan digunakan proses deteksi *outlier*. Pendeteksian *outlier*/pencilan merupakan deteksi pada data yang polanya tidak mengikuti model umum, atau nilai kesalahannya ada di nilai tiga kali dari nilai simpangan bakunya atau dari nilai rata-rata sisaannya yang tidak mendekati sama sekali [2].

Dalam mendeteksi *outlier* ada beberapa metode yang digunakan yaitu pencarian nilai titik *influence* dengan metode *DfFit* dan *DfBeta*. Namun, dalam menghitung nilai *DfFit* dan *DfBeta* ada perhitungan awal yang harus dilakukan yaitu menghitung nilai dari titik *leverage*. Titik *leverage* terbagi menjadi dua yaitu titik *leverage* baik (*good leverage point*) dan titik *leverage* buruk (*bad leverage point*). *Leverage point* yang baik adalah *outlier* yang disebabkan oleh variabel prediktor saja, sedangkan *leverage point* yang buruk adalah *outlier* yang disebabkan oleh variabel respon dan prediktor [3]. Nilai h_{ii} melampaui *cutoff*-nya maka data tersebut terdeteksi sebagai *outlier* [4].

Metode lain yang digunakan dalam deteksi *outlier* adalah metode titik *infuence* ada dua yang digunakan yaitu metode *global first effect* (*DfFit* dan jarak *Cook's*) yang

menjelaskan tentang ketika karakteristik global model regresi observasi ke-i dipengaruhi. Sedangkan yang kedua adalah *direct effect* (*DfBeta*) yang menggambarkan bagaimana kasus ke-i mempengaruhi masing-masing koefisien regresi [5]. Metode *DfFit* digunakan untuk mengukur perubahan dalam nilai yang diprediksi ketika suatu pengamatan tertentu Ketika dimasukkan dandikeluarkan dalam estimasi. Sedangkan, metode *DfBeta* digunakan untuk mengukur perbedaan antara koefisien suatu prediktor tertentu ketika suatu pengamatan tertentu Ketika dikeluarkan dan dimasukkan ke dalam regresi [6].

Salah satu data yang menarik untuk dideteksi pencilannya adalah data longitudinal. Data longitudinal merupakan penggabungan data suatu observasi (*cross-section)* dengan data deret waktu (*time series)* [7]. Data longitudinal mampu membedakan keragaman respon yang terjadi karena pengukuran yang dilakukan berulang-ulang pada suatu subjek dengan keragaman yang terjadi oleh perbedaan antar subjek. Menurut Wu dan Zhang (2006), ada tiga keuntungan studi longitudinal dibandingkan dengan studi *cross-section*. Studi longitudinal memiliki kekuatan yang lebih besar pada sejumlah observasi yang tetap dengan kata lain, studi longitudinal membutuhkan lebih sedikit subjek untuk menghasilkan kekuatan uji statistik yang sama. Pada banyaknya subjek yang sama, hasil galat yang diukur dapat dikurangi untuk mengurangi efek perlakuan. Data longitudinal dapat memberikan penjelasan mengenai individu yang berubah [8].

Pendeteksian *outlier* tersebut, dapat dilakukan pada berbagai bidang penelitian, misalnya bidang pendidikan, kesehatan, ekonomi, serta yang lainnya. Maka dari itu, berikut merupakan contoh penerapannya melalui bidang kesehatan berkaitan dengan data kesehatan di Jawa Barat pada tahun 2021-2022. Berdasarkan uraian diatas, maka peneliti menggunakan studi kasus yang diperoleh dari *website* Badan Pusat Statistik (BPS) dan Open Data Jawa Barat yang menunjukkan data kesehatan di Jawa Barat pada rentang waktu 2021-2022 di sejumlah 27 Kab/kota yang ada di Jawa Barat. Data yang diambil meupakan data kasus stunting. Menurut *World Health Organization* (WHO) pada tahun 2025, stunting pada balita adalah gangguan pertumbuhan dan perkembangan anak yang disebabkan oleh kekurangan gizi yang berkelanjutan dan

tidak adekuat, infeksi berulang, dan tinggi badan yang kurang dari standar atau kriteria. WHO juga menyatakan bahwa stunting pada balita adalah pendek atau sangat pendek jika tinggi badannya kurang dari -2 standar deviasi (SD) berdasarkan kurva pertumbuhan yang dikemukakan oleh WHO. Berdasarkan hasil SSGI 2021 prevalensi stunting di Jawa Barat menempati angka sebesar 24,5%, terjadi penurunan sebesar 1,35% untuk angka rata-rata stunting di Jawa Barat pada tiga tahun terakhir. Salah satu variabel yang diambil dalam penelitian ini adalah wasting dan stunting yang masing-masing saling berkaitan. Anak-anak yang mengalami *wasting* berisiko lebih tinggi untuk menjadi stunting di kemudian hari. Ini menunjukkan bahwa wasting dan stunting saling terkait, di mana *wasting* dapat menjadi faktor risiko untuk perkembangan stunting. Intervensi dini untuk mencegah *wasting* dapat berdampak signifikan dalam mengurangi prevalensi stunting, yang merupakan indikator malnutrisi kronis yang berdampak pada perkembangan anak. Kasus stunting yang merupakan bentuk dari data longitudinal terjadi di Gambia yaitu meneliti mengenai kasus stunting dan wasting pada tahun 1976 sampai 2016 [9].

Data yang diperoleh merupakan data longitudinal dari 27 Kab/kota di Jawa Barat selama rentang waktu 2021-2022 yang kemudian nantinya akan dihitung nilai titik *leverage*-nya untuk menentukan apakah data tersebut termasuk *outlier* atau tidak. Lalu dicari nilai *internally studentized resiuals* untuk menentukan *outlier* pada tahap kedua. Selanjutnya dicari nilai titik pengaruhnya dengan dua metode yaitu metode *DfFit* dan *DfBeta*. Sehingga penulis akan membuat skripsi yang berjudul "Pendeteksian Outlier pada Data Longitudinal dengan Metode *DfFit* dan *DfBeta* untuk Data Kesehatan di Jawa Barat."

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah disampaikan di atas, maka masalah yang akan dirumuskan pada penelitian ini, yaitu :

1. Dalam suatu data terdapat data yang tidak mengikuti pola umum model biasanya residual (*error*-nya) tiga kali dari nilai standar deviasi-nya atau lebih

jauh dari nilai rata-rata residualnya ini mempengaruhi estimasi yang tidak akurat dari parameter (*slopes* dan *intercept*).

2. Keberadaan *outlier* sering kali dapat merusak estimasi parameter dalam regresi linear biasa (*Least Squares*) sehingga tidak memberikan hasil yang memadai karena data yang tidak memenuhi asumsi-asumsi klasik dari regresi linear.

1.3 Batasan Masalah

Dalam perumusan masalah, adapun beberapa hal yang menjadi batasan dari masalah tersebut diantaranya adalah :

1. Metode pendeteksian *outlier* yang digunakan adalah pencarian titik *influence* dengan metode *DfFit* dan *DfBeta*.
2. Pendeteksian *outlier* dilakukan pada jenis data longitudinal.
3. Model estimasi yang digunakan yaitu model estimasi OLS.
4. Data yang digunakan dalam penelitian ini adalah data kasus stunting di Jawa Barat tahun 2021-2022.

1.4 Tujuan Penelitian

Berdasarkan latar belakang dan rumusan masalah yang telah disampaikan, maka terdapat beberapa tujuan dari penelitian ini, diantaranya:

1. Mendeteksi *outlier* pada data keseluruhan dan mengidentifikasi data yang termasuk *outlier* dengan metode *DfFit* dan *DfBeta* lalu menghilangkan data tersebut, melakukan estimasi parameter OLS untuk membandingkan tingkat penurunan nilai *error*-nya sehingga didapat metode deteksi *outlier* terbaik.
2. Menghilangkan *outlier* dari data keseluruhan, melakukan proses estimasi OLS pada masing-masing metode deteksi *outlier* dibandingkan hasil *residual error*, koefisien determinasi, *Adjusted* R^2, dan juga RMSE untuk mendapatkan model estimasi OLS yang akurat.

1.5 Metode Penelitian

1. Metode yang digunakan dalam penelitian ini diperoleh dari berbagai sumber buku, artikel, jurnal, skripsi, dan thesis yang berkaitan dengan topik penelitian.
2. *Software* yang digunakan untuk melakukan proses perhitungan pada setiap metode adalah *software* R Studio.

1.6 Sistematika Penulisan

Berdasarkan cara penulisannya, penelitian ini menggunakan sistematika penulisan yang terdiri dari lima bab, dan setiap bab terdiri dari beberapa sub bab.

BAB I PENDAHULUAN

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

BAB II LANDASAN TEORI

Bab ini berisi mengenai teori-teori yang berhubungan dengan topik penelitian, dan secara umum mencakup seluruh materi yang berhubungan dengan pendeteksian outlier pada data longitudinal yang terdiri atas Data Longitudinal, Model Analisis Regresi, Regresi Data Longitudinal, Estimasi Parameter *Ordinary Least Square*, Uji Asumsi Model Regresi Data Longitudinal, Koefisien Determinasi, Deteksi *Outlier,* dan R Studio.

BAB III PENDETEKSIAN *OUTLIER* PADA DATA LONGITUDINAL DENGAN METODE *DFFIT* DAN *DFBETA* UNTUK DATA KESEHATAN DI JAWA BARAT

Bab ini berisi mengenai bahasan utama pada penelitian ini, yang meliputi cara pendeteksian outlier menggunakan metode *DfFit* dan *DfBeta* pada data longitudinal dan estimasi parameternya.

BAB IV STUDI KASUS DAN ANALISA

Bab ini berisi studi kasus dan analisa data tentang bahasan utama yang telah dijelaskan pada bab iii yang diaplikasikan dalam data Kesehatan di Jawa Barat untuk Kasus Stunting pada tahun 2021-2022.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan dan saran yang bisa diajukan pada penelitian ini.