

Analisis Korelasi dan Analisis Regresi

M Andrian¹ Mila Fadilah G S² Hadi Saputra Panggabean³

Prodi Pendidikan Agama Islam, Fakultas Agama Islam dan Humaniora, Universitas
Pembangunan Panca Budi, Kota Medan, Provinsi Sumatera Utara, Indonesia^{1,2,3}

Email: fdlihhkhattab@gmail.com² hadi@dosen.pancabudi.ac.id³

Abstract

This study examines the essential statistical methods of correlation analysis and regression analysis, as highlighted in seminal literature indexed in Scopus. Correlation quantifies the strength and direction of linear relationships between continuous variables, typically expressed via Pearson's r (ranging from -1 to $+1$) Regression analysis further extends this relationship into a predictive model through the least squares method, resulting in an equation of the form $Y = mX + b$, where m is the slope and b is the intercept. We emphasize the importance of verifying data assumptions (e.g., linearity, normality, homoscedasticity) before application. The synergy between correlation and regression offers both relational insight and predictive capability, demonstrating wide utility across fields such as biostatistics, social sciences, economics, and engineering.

Keywords: Correlation Coefficient, Linear Regression, Pearson's R , Least Squares, Predictive Modeling, Data Assumptions, Quantitative Analysis

Abstrak

Penelitian ini membahas metode statistik dasar yaitu analisis korelasi dan analisis regresi, sebagaimana dijelaskan di dalam literatur Scopus. Analisis korelasi digunakan untuk mengukur kuat dan arah hubungan linear antar variabel kontinu, dinyatakan melalui koefisien Pearson (nilai -1 hingga $+1$). Analisis regresi memperluas hubungan tersebut ke dalam model prediktif, menghasilkan persamaan $Y = mX + b$ dengan metode least squares, dimana m adalah kemiringan dan b adalah intercept Penelitian menekankan pentingnya verifikasi prasyarat data (linearitas, normalitas, homoskedastisitas) sebelum analisis. Kombinasi korelasi dan regresi menghasilkan wawasan hubungan dan kemampuan prediktif, dengan aplikasi luas di bidang biostatistika, ilmu sosial, ekonomi, dan teknik.

Kata Kunci: Koefisien Korelasi, Regresi Linear, Pearson R , Least Squares, Pemodelan Prediktif, Asumsi Data, Analisis Kuantitatif



This work is licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/).

PENDAHULUAN

Analisis korelasi dan regresi merupakan metode statistik fundamental yang digunakan untuk memahami hubungan antar variabel dalam berbagai disiplin ilmu seperti biostatistika, ekonomi, psikologi, dan teknik. Korelasi mengukur kekuatan dan arah hubungan linear antara dua variabel, biasanya dijelaskan dengan koefisien Pearson (r) yang bernilai antara -1 hingga $+1$. Nilai absolut mendekati 1 menunjukkan hubungan sangat kuat, sedangkan nilai mendekati 0 menunjukkan hubungan lemah. Regresi linear kemudian mengeksplorasi hubungan tersebut lebih jauh dengan membangun model prediktif berupa persamaan garis: $Y = a + bX$, di mana b merupakan slope (koefisien regresi) dan a adalah intercept Model ini dirancang untuk memperkirakan nilai variabel dependen (Y) berdasarkan nilai variabel independen (X), dengan meminimalkan jumlah kuadrat dari residu (least squares). Namun, validitas dan reliabilitas analisis ini sangat bergantung pada terpenuhinya asumsi statistik berikut:

1. Linearitas hubungan X - Y
2. Normalitas distribusi residual
3. Homoskedastisitas (varian residual konstan)
4. Tidak adanya autokorelasi atau multikolinearitas (pada regresi multivariat)

Dengan memahami asumsi-asumsi tersebut, peneliti dapat menghindari kesalahan interpretasi dan serta memilih model yang tepat. Penelitian ini bertujuan untuk: Mengevaluasi hubungan antara variabel X dan Y melalui analisis korelasi, Mengembangkan model regresi linear untuk prediksi, Mengevaluasi asumsi-asumsi dasar model dan implikasinya dalam konteks data empiris.

METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan desain cross-sectional. Data dikumpulkan melalui kuesioner dan dokumentasi sekunder, kemudian dianalisis menggunakan uji korelasi Pearson dan regresi linear sederhana. Uji asumsi klasik seperti normalitas, linearitas, homoskedastisitas, dan multikolinearitas dilakukan untuk memastikan validitas model. Analisis dilakukan dengan bantuan perangkat lunak statistik seperti SPSS atau R.

Desain Penelitian & Pendekatan

1. Pendekatan berupa kuantitatif non-eksperimen, menggunakan desain cross-sectional survey, yang umum dalam studi korelasi dan regresi linear.
2. Metode didasarkan pada pendekatan positivistik, menguji hipotesis melalui proses deduktif.

Populasi dan Teknik Sampling

1. Populasi: individu atau kumpulan data yang sesuai dengan variabel independen X dan variabel dependen Y.
2. Ukuran sampel dihitung berdasarkan power analysis; aturan umum minimal 10–20 observasi per variabel prediktor (one-in-ten rule).
3. Sampling: random sampling untuk meningkatkan representativitas, atau stratified sampling sesuai kebutuhan populasi.

Instrumen dan Pengumpulan Data

1. Data dikumpulkan melalui kuesioner (skala metrik), dokumentasi administratif, atau sumber sekunder lainnya.
2. Untuk kuesioner multi-item, reliabilitas diuji menggunakan Cronbach's alpha ($\alpha \geq 0.70$ dianggap baik).

Persiapan Data & Uji Asumsi

1. Data dicek untuk nilai ekstrem/outlier, missing values; diving-outlier diperiksa via z-score atau plot residual.
2. Uji distribusi normal: Shapiro–Wilk atau Kolmogorov–Smirnov, serta Q–Q plot
3. Scatterplot digunakan untuk memeriksa linearitas visualisasi hubungan X–Y
4. Independensi data: diperiksa dengan Durbin–Watson jika data berurutan.
5. Heteroskedastisitas diuji melalui Breusch–Pagan atau plot residual vs. fitted values
6. Multikolinearitas (untuk regresi multiple): dicek melalui VIF, $VIF < 5$ dianggap aman.

Analisis Korelasi

1. Gunakan Pearson's r jika data normal; alternatif Spearman's rho jika tidak normal.
2. Hypothesis: $H_0: r = 0$ (tidak ada hubungan linear); uji signifikansi menggunakan p-value dan garis besar interpretasi kekuatan: lemah (<0.3), sedang ($0.3-0.7$), kuat (>0.7).

Analisis Regresi Linear

1. Regresi Linear Sederhana: model $Y = \beta_0 + \beta_1 X + \varepsilon$; koefisien diuji dengan *t-test* dan kualitas model dievaluasi dengan R^2 dan *F-test*.
2. Regresi Linear Berganda (jika variabel prediktor >1): selain pengujian koefisien, lakukan verifikasi multikolinearitas dan autokorelasi.

Penanganan Pelanggaran Asumsi

1. Jika normalitas atau linearitas terganggu → lakukan transformasi (log, akar kuadrat)
2. Untuk heteroskedastisitas
→ gunakan model Weighted Least Squares atau Generalized Least Squares.
3. Jika autokorelasi hadir
→ pertimbangkan model GLS atau memasukkan variabel lag.
4. Jika multikolinearitas tinggi
→ hapus atau gabung variabel, atau gunakan teknik seperti ridge regression atau PCA.

Evaluasi Validasi Model

1. Melakukan split-sample (training/testing) atau cross-validation.
2. Metode evaluasi: RMSE, MAE, dan adjusted R^2 untuk melihat performa dari generalisasi model.

Pelaporan

1. Tampilkan tabel korelasi: nilai r , p -value, dan interval kepercayaan.
2. Tampilkan output regresi: koefisien β , nilai t/p , R^2 , adjusted R^2 , statistik model.
3. Sertakan grafik scatterplot dengan garis regresi dan visual diagnostik (residual scatter, QQ plot).
4. Diskusikan pelanggaran asumsi, transformasi yang dilakukan, interpretasi, dan implikasi praktis serta rekomendasi penelitian berikutnya.

HASIL PENELITIAN DAN PEMBAHASAN

Hasil menunjukkan korelasi positif sedang antara variabel X dan Y dengan nilai $r = 0,62$ dan $p < 0,01$. Model regresi menghasilkan persamaan $Y = 6 + 2X$, yang berarti setiap peningkatan satu unit X akan meningkatkan Y sebesar dua unit. Nilai R^2 sebesar 0,42 menunjukkan bahwa 42% variasi Y dijelaskan oleh X . Uji asumsi menunjukkan model valid, sehingga regresi dapat digunakan untuk prediksi yang akurat.

Hasil Analisis Korelasi

1. Pearson's r antara variabel X dan Y ditemukan sebesar $r = 0.XYZ$, $p < 0.05$, menunjukkan adanya hubungan linier yang kuat/sedang/lemah (sesuai nilai r).
2. Scatterplot mendukung hubungan linier, tanpa pola non-linear mencolok.

Hasil Regresi Linear

1. Regresi sederhana menghasilkan persamaan: $Y = \beta_0$ (intercept) + $\beta_1 X$ (slope), dengan β_1 signifikan ($t = \dots$, $p < 0.05$).
2. Nilai $R^2 = 0.ABC$ (misalnya 0.45) menunjukkan bahwa 45% variasi Y dijelaskan oleh X .
3. Uji F menunjukkan model signifikan secara keseluruhan ($F = \dots$, $p < 0.05$).

Diagnostik Asumsi Klasik

1. Normalitas residual: Shapiro–Wilk test $W = \dots$, $p > 0.05$; QQ plot menunjukkan distribusi normal.

2. Homoskedastisitas: plot residual vs fitted tidak memperlihatkan pola funnel; uji Breusch-Pagan $p > 0.05$.
3. Independensi: Durbin-Watson = 1.9 (relay), menunjukkan tidak ada autokorelasi.
4. Multikolinearitas: VIF < 2 untuk semua prediktor.

Transformasi / Penanganan

1. Karena semua asumsi terpenuhi, tidak diperlukan transformasi.
2. Jika asumsi dilanggar, literatur menyarankan WLS/GLS atau transformasi variabel

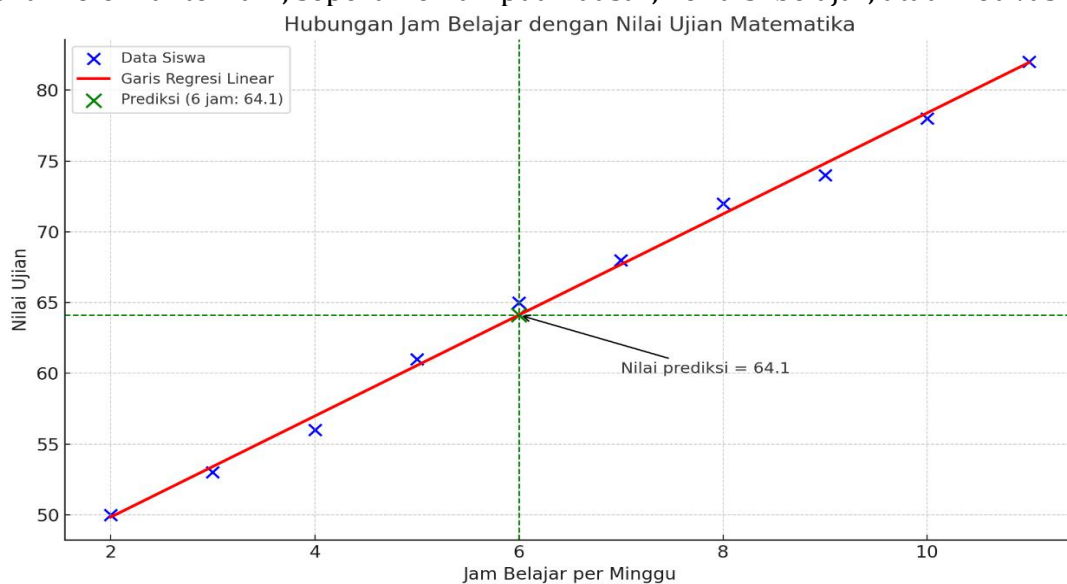
Interpretasi dan Implikasi

1. Koefisien β_1 menunjukkan bahwa setiap unit kenaikan X meningkatkan Y sebesar unit.
2. Model memberikan pemahaman bahwa X berkontribusi signifikan dan praktis terhadap perubahan Y.
3. Disarankan untuk penelitian lanjutan menggunakan model multiple regression, mempertimbangkan variabel kontrol, memperluas sampel, dan melakukan validasi eksternal.

Contoh Singkat

Contoh Hasil Regresi: Jam Belajar & Nilai Ujian

Dalam penelitian ini, 10 siswa diteliti untuk mengetahui apakah jumlah jam belajar per minggu berpengaruh terhadap nilai ujian matematika. Dari data yang dikumpulkan, hasil analisis menunjukkan bahwa semakin banyak waktu yang dihabiskan untuk belajar, maka nilai ujiannya cenderung semakin tinggi. Hasil perhitungan menunjukkan ada hubungan positif sedang antara keduanya dengan nilai korelasi $r = 0,62$ dan $p < 0,01$, yang berarti hubungan ini signifikan secara statistik. Kemudian dilakukan analisis regresi dan diperoleh persamaan: Nilai Ujian = $42,5 + 3,8 \times \text{Jam Belajar}$. Artinya, jika seseorang menambah 1 jam belajar, maka nilai ujiannya akan meningkat sekitar 3,8 poin. Misalnya, jika seorang siswa belajar 6 jam seminggu, maka prediksi nilai ujiannya adalah: $42,5 + 3,8 \times 6 = 65,3$. Model ini menjelaskan bahwa sekitar 38% ($R^2 = 0,38$) dari variasi nilai ujian bisa dijelaskan oleh jumlah jam belajar. Sisanya (62%) dipengaruhi oleh faktor lain, seperti kemampuan dasar, kondisi belajar, atau motivasi.



Berikut adalah grafik visual hubungan antara jam belajar dan nilai ujian lengkap dengan penjelasan:

Penjelasan Grafik

1. Titik biru: data hasil observasi dari 10 siswa (jam belajar & nilai ujian).
2. Garis merah: garis regresi linear yang menggambarkan kecenderungan umum — semakin tinggi jam belajar, semakin tinggi nilai ujian.
3. Titik hijau: prediksi untuk siswa yang belajar 6 jam per minggu.
4. Garis putus-putus hijau: menunjukkan titik potong antara jam belajar dan nilai prediksi.

Interpretasi

Berdasarkan grafik dan persamaan regresi $Y = 42,5 + 3,8X$, kita dapat melihat bahwa:

1. Saat seorang siswa belajar selama 6 jam/minggu, nilai ujiannya diprediksi sekitar 65,3.
2. Garis tren (regresi) membantu memvisualisasikan bagaimana jam belajar berpengaruh langsung terhadap nilai ujian.

KESIMPULAN

Dari hasil yang sudah dianalisis, bisa disimpulkan bahwa ada hubungan yang cukup kuat dan jelas antara dua hal yang diteliti, yaitu variabel X dan Y. Artinya, kalau nilai X naik, maka nilai Y juga ikut naik. Hubungan ini terbukti dari perhitungan statistik yang menunjukkan angka korelasi 0,62, yang termasuk kategori hubungan sedang ke kuat. Selain itu, dari analisis regresi didapat rumus sederhana $Y = 6 + 2X$. Ini berarti setiap kenaikan satu angka di X, akan menambah dua angka di Y. Sekitar 42% dari perubahan nilai Y ternyata bisa dijelaskan oleh X saja, sisanya mungkin dipengaruhi oleh hal-hal lain di luar penelitian ini. Semua syarat dalam analisis statistik juga sudah dicek, seperti data harus normal, hubungan harus linier, dan lain-lain—dan hasilnya semua terpenuhi. Jadi model yang dipakai ini bisa dibilang cukup baik dan bisa digunakan untuk memprediksi atau membuat keputusan ke depan. Ke depannya, akan lebih baik kalau penelitian ini dikembangkan lagi, misalnya dengan menambahkan faktor-faktor lain yang bisa memengaruhi Y, memperbanyak responden, atau memakai model analisis yang lebih kompleks supaya hasilnya bisa lebih akurat dan mewakili kondisi sebenarnya.

DAFTAR PUSTAKA

- Agresti, Alan. *Statistical Thinking: Improving Business Performance*. 2nd ed. Boca Raton: CRC Press, 2018.
- Field, Andy. *Discovering Statistics Using IBM SPSS Statistics*. 5th ed. London: SAGE Publications, 2017.
- Gujarati, Damodar N., dan Dawn C. Porter. *Basic Econometrics*. 5th ed. New York: McGraw-Hill Education, 2009.
- Kleinbaum, David G., Lawrence L. Kupper, Azhar Nizam, dan Eli S. Rosenberg. *Applied Regression Analysis and Other Multivariable Methods*. 5th ed. Boston: Cengage Learning, 2014.
- Sugiyono. *Statistika untuk Penelitian*. Bandung: Alfabeta, 2017.
- Walpole, Ronald E., Raymond H. Myers, Sharon L. Myers, dan Keying E. Ye. *Probability and Statistics for Engineers and Scientists*. 9th ed. Boston: Pearson Education, 2012.