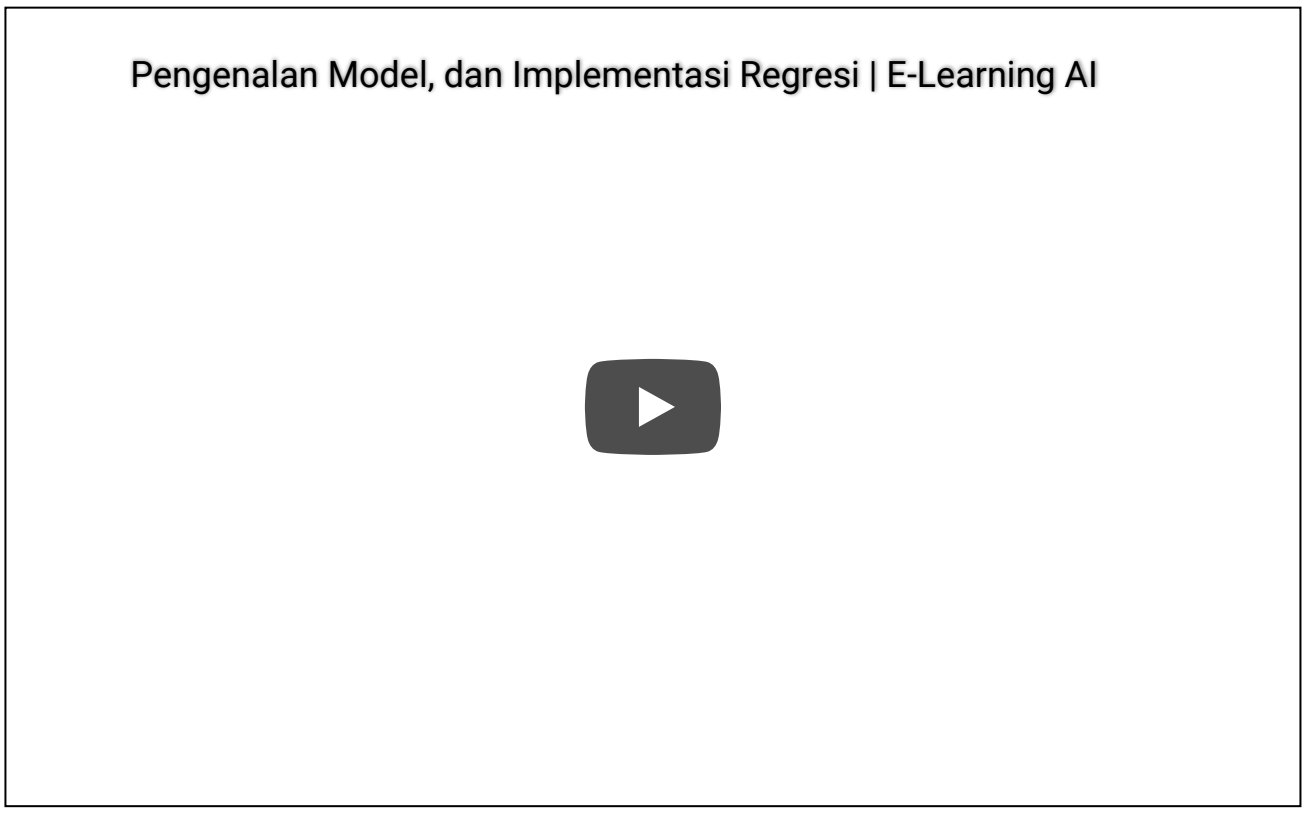
Dasar-Dasar Statistika Ilmu Komputer : Regresi

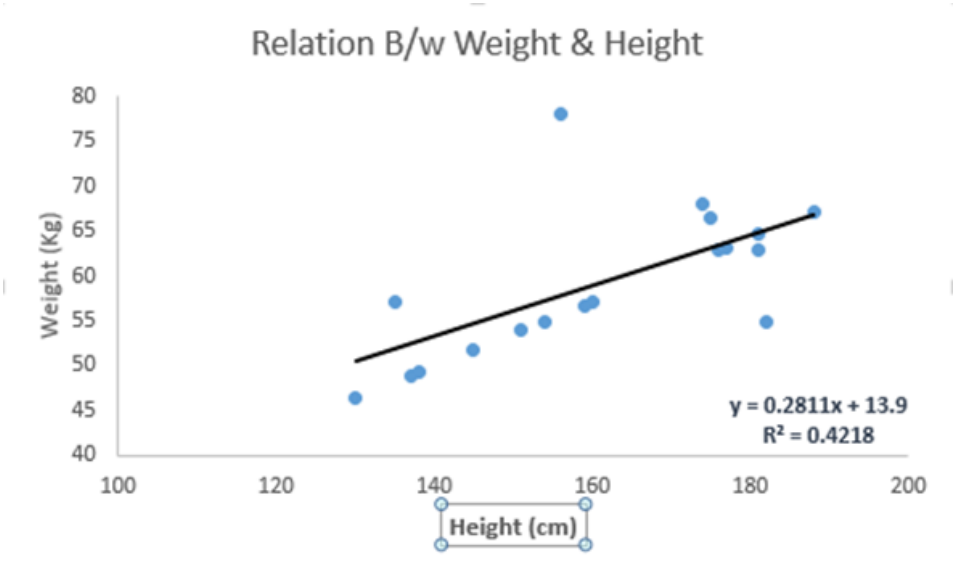


Dasar teori

Introduction
Regresi linier dan logistik biasanya merupakan algoritma pertama yang dipelajari seorang Data Science. Karena popularitasnya, banyak analis bahkan akhirnya berpikir bahwa itu adalah satu-satunya bentuk regresi. Sebenarnya ada banyak sekali bentuk regresi, yang dapat digunakan sesuai dengan keperluan dan kondisi khusus di mana mereka paling cocok untuk diterapkan.

Linear Regression
Regresi Linier adalah salah satu teknik pemodelan yang paling dikenal. Regresi linier biasanya topik pertama yang dipilih orang saat mempelajari pemodelan prediktif. Dalam teknik ini, variabel terikat bersifat kontinu, variabel bebas dapat kontinu atau diskrit, dan sifat garis regresinya linier. Regresi Linier menetapkan hubungan antara variabel terikat (Y) dan satu atau lebih variabel bebas (X) menggunakan garis lurus yang paling sesuai (juga dikenal sebagai garis regresi).

Hal ini diwakili oleh persamaan Y = a + bx, di mana a adalah intersep dan b adalah kemiringan garis. Persamaan ini dapat digunakan untuk memprediksi nilai variabel target berdasarkan variabel prediktor yang diberikan.



Perbedaan regresi linier sederhana dan regresi linier berganda adalah, regresi linier berganda memiliki (>1) variabel bebas, sedangkan regresi linier sederhana hanya memiliki 1 variabel bebas. Sekarang, pertanyaannya adalah “Bagaimana kita mendapatkan garis yang paling cocok?”.

Bagaimana cara mendapatkan garis yang paling sesuai (Nilai a dan b)?
Pertanyaan ini dapat dengan mudah diselesaikan dengan Metode Kuadrat Terkecil. Ini adalah metode yang paling umum digunakan untuk memasang garis regresi. Dengan menghitung garis yang paling cocok untuk data yang diamati dengan meminimalkan jumlah kuadrat deviasi vertikal dari setiap titik data ke garis. Karena deviasi dikuadratkan terlebih dahulu, ketika ditambahkan, tidak ada pembatalan antara nilai positif dan negatif. Kita dapat mengevaluasi kinerja model menggunakan metrik R-Square.

Poin Penting:
- Harus ada hubungan linier antara variabel bebas dan variabel terikat
- Dalam regresi linier berganda terdapat multikolinearitas, autokorelasi, heteroskedastisitas.
- Regresi Linier sangat sensitif terhadap Outlier. Ini dapat sangat mempengaruhi garis regresi dan nilai perkiraan (prediksi).
- Multikolinearitas dapat meningkatkan varians dari estimasi koefisien dan membuat estimasi sangat sensitif terhadap perubahan kecil dalam model. Hasilnya adalah estimasi koefisien tidak stabil

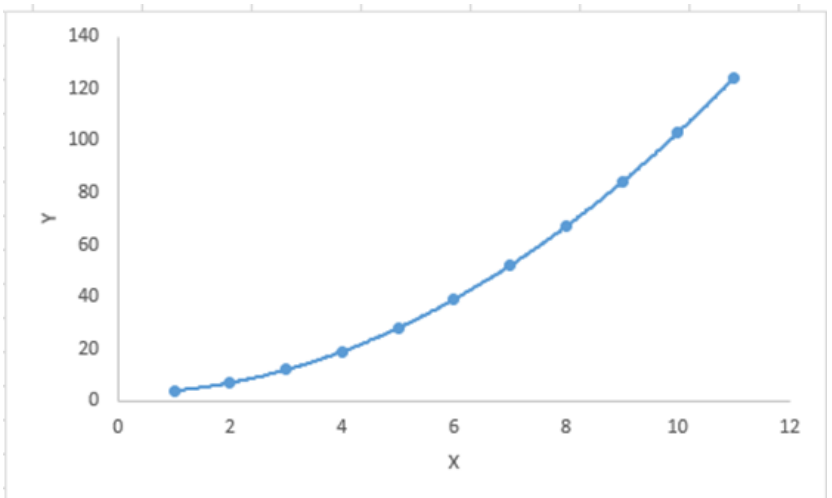
Logistic Regression
Regresi logistik digunakan untuk mencari peluang event=Success dan event=Failure. Kita harus menggunakan regresi logistik ketika variabel dependen adalah biner (0/1, Benar/ Salah, Ya/ Tidak). Di sini nilai Y berkisar dari 0 hingga 1 dan dapat diwakili oleh persamaan berikut.

Peluang = p/ (1-p) = peluang terjadinya peristiwa / peluang terjadinya bukan peristiwa
ln(peluang) = ln(p/(1-p))
logit(p) = ln(p/(1-p)) = b0+b1X1+b2X2+b3X3+.... +bkXk

Pada penjelasan diatas, p adalah probabilitas kehadiran karakteristik yang diinginkan. Kemudian mungkin ada pertanyaan “mengapa kita menggunakan persamaan log in?”.
Karena kita bekerja di sini dengan distribusi binomial (variabel dependen), kita perlu memilih link function yang paling cocok untuk distribusi ini. Dan, itu adalah fungsi logit. Dalam persamaan di atas, parameter dipilih untuk memaksimalkan kemungkinan mengamati nilai sampel daripada meminimalkan jumlah kesalahan kuadrat (seperti dalam regresi biasa).

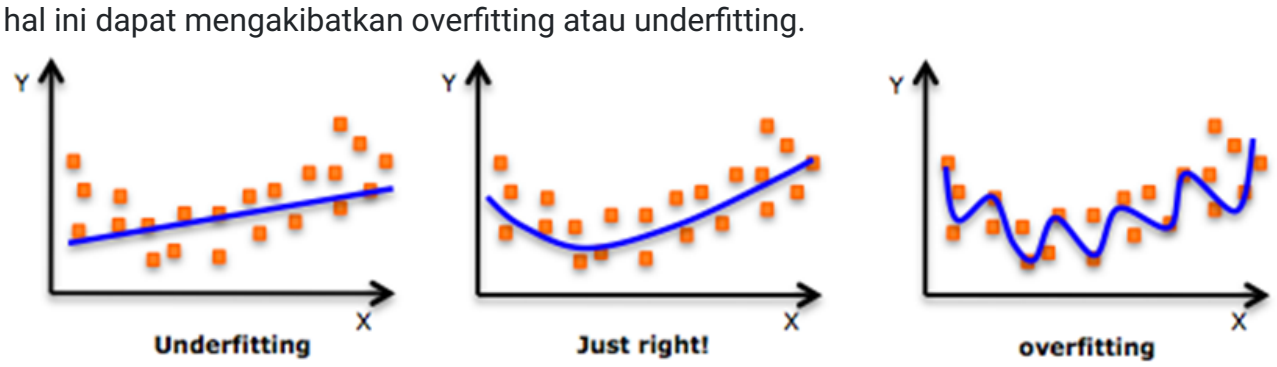
Poin Penting:
- Regresi logistik banyak digunakan untuk masalah klasifikasi
- Regresi logistik tidak membutuhkan hubungan linier antara variabel dependen dan independen. Ini dapat menangani berbagai jenis hubungan karena menerapkan transformasi log non-linier ke rasio peluang yang diprediksi
- Untuk menghindari over fitting dan under fitting, kita harus memasukkan semua variabel yang signifikan. Pendekatan yang baik untuk memastikan praktik ini adalah dengan menggunakan metode langkah bijaksana untuk memperkirakan regresi logistik
- Regresi logistik membutuhkan ukuran sampel yang besar karena perkiraan kemungkinan maksimum kurang kuat pada ukuran sampel rendah daripada kuadrat terkecil biasa
- Variabel bebas tidak boleh dikorelasikan satu sama lain, artinya tidak ada multikolinieritas. Namun, memiliki pilihan untuk memasukkan efek interaksi variabel kategori dalam analisis dan model.
- Jika nilai variabel terikatnya ordinal, maka disebut regresi logistik Ordinal
- Jika variabel terikatnya adalah multi kelas maka disebut dengan regresi Logistik Multinomial.

Polynomial Regression
Regresi Polinomial adalah persamaan regresi yang memiliki pangkat variabel bebas lebih dari 1. Persamaan di bawah ini merupakan persamaan polinomial: Y = a + b*x^2



Dalam teknik regresi ini, garis yang paling sesuai bukanlah garis lurus. Ini lebih merupakan kurva yang cocok dengan titik-titik data.

Poin Penting:
- Jika kitamelakukan penyesuaian polinomial dengan derajat yang lebih tinggi untuk mendapatkan kesalahan yang lebih rendah, hal ini dapat mengakibatkan overfitting atau underfitting.



Stepwise Regression

Bentuk regresi ini digunakan ketika kita berurusan dengan beberapa variabel independen. Dalam teknik ini, pemilihan variabel bebas dilakukan dengan bantuan proses otomatis, yang tidak melibatkan campur tangan manusia. Hal tersebut dilakukan dengan mengamati nilai statistik seperti R-square, t-stats dan metrik AIC untuk membedakan variabel yang signifikan. Stepwise Regression pada dasarnya cocok dengan model regresi dengan menambahkan/mendrop kovariat satu per satu berdasarkan kriteria yang ditentukan.

Beberapa metode Stepwise Regression yang paling umum digunakan seperti di bawah ini:
- Regresi bertahap standar melakukan dua hal. Ini menambah dan menghapus prediktor sesuai kebutuhan untuk setiap langkah.
- Pemilihan ke depan dimulai dengan prediktor paling signifikan dalam model dan menambahkan variabel untuk setiap langkah.
- Tujuan dari teknik pemodelan ini adalah untuk memaksimalkan daya prediksi dengan jumlah variabel prediktor yang minimum.

Ridge Regression

Regresi Ridge adalah teknik yang digunakan ketika data mengalami multikolinearitas. Dalam multikolinearitas, meskipun perkiraan kuadrat terkecil (OLS) tidak bias, variansnya besar yang menyimpang nilai yang diamati jauh dari nilai sebenarnya dengan menambahkan derajat bias pada

Dalam persamaan linier, kesalahan prediksi dapat didekomposisi menjadi dua sub komponen. Pertama karena bias dan kedua karena varians. Kesalahan prediksi dapat terjadi karena salah satu dari dua atau kedua komponen ini. Kita akan membahas tentang kesalahan yang disebabkan karena varians. Regresi ridge menyelesaikan masalah multikolinearitas melalui parameter penyusutan (lambda). Perhatikan persamaan di bawah ini.

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Penalty}}$$

Dalam persamaan ini, kita memiliki dua komponen. Yang pertama adalah suku kuadrat terkecil dan yang lainnya adalah lambda penjumlahan 2 (beta-kuadrat) yang mana adalah koefisiennya. Ini ditambahkan ke suku kuadrat terkecil untuk mengecilkan parameter agar memiliki varians yang sangat rendah.

Poin Penting:
- Asumsi regresi ini sama dengan regresi kuadrat terkecil kecuali normalitas tidak diasumsikan
- Ridge regresi mengecilkan nilai koefisien tetapi tidak mencapai nol, yang menunjukkan tidak ada fitur pemilihan fitur

Lasso Regression

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}$$

Mirip dengan Regresi Ridge, Lasso (Least Absolute Shrinkage and Selection Operator) juga menilai ukuran absolut dari koefisien regresi. Selain itu, mampu mengurangi variabilitas dan meningkatkan akurasi model regresi linier. Regresi Lasso berbeda dari regresi ridge karena menggunakan nilai absolut dalam fungsi penalti, bukan kuadrat. Hal ini secara ekuivalen membatasi jumlah nilai absolut dari estimasi dan menyebabkan beberapa estimasi parameter menjadi benar-benar nol.

Poin Penting:
- Asumsi regresi lasso sama dengan regresi kuadrat terkecil kecuali normalitasnya tidak diasumsikan
- Regresi Lasso menyusutkan koefisien menjadi nol, yang tentunya membantu dalam pemilihan fitur
- Lasso adalah metode regularisasi dan menggunakan regularisasi l1
- Jika kelompok prediktor sangat berkorelasi, lasso hanya mengambil salah satu dari mereka dan mengecilkan yang lain menjadi nol

ElasticNet Regression

ElasticNet adalah hibrid dari teknik Lasso dan Ridge Regression. Hal ini dilatih dengan L1 dan L2 sebelumnya sebagai regularizer. Elastic-net berguna ketika ada beberapa fitur yang berkorelasi. Lasso kemungkinan akan memilih salah satunya secara acak, sementara elastic-net kemungkinan akan memilih keduanya.

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1).$$

Keuntungan praktis dari pertukaran antara Lasso dan Ridge adalah,memungkinkan Elastic-Net untuk mewarisi beberapa stabilitas Ridge di bawah rotasi.

Poin Penting:
- Mendorong group effect dalam kasus variabel yang sangat berkorelasi
- Tidak ada batasan jumlah variabel yang dipilih

- Dapat mengalami double shrinkage

Quiz