# Regression

Regresi merupakan salah satu teknik untuk meramalkan data di masa yang akan datang. Lebih mudahnya mari kita lihat kasus di mana perusahaan mendata pegawainya berdasarkan berapa lama mereka bekerja dan berapa besar gaji mereka saat ini (dalam euro per tahun). Ilustrasinya tampak pada tabel di bawah ini:

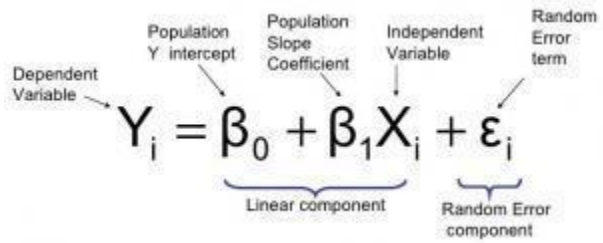
| Tahun_bekerja | Gaji |
|---|---|
| 1.1 | 39,343 |
| 1.3 | 46,205 |
| 1.5 | 37,731 |
| 2.0 | 43,525 |
| 2.2 | 39,891 |
| 2.9 | 56,642 |
| 3.0 | 60,150 |
| 3.2 | 54,445 |
| 3.2 | 64,445 |
| 3.7 | 57,189 |
| 3.9 | 63,218 |
| 4.0 | 55,794 |
| 4.0 | 56,957 |
| 4.1 | 57,081 |
| 4.5 | 61,111 |
| 4.9 | 67,938 |
| 5.1 | 66,029 |
| 5.3 | 83,088 |
| 5.9 | 81,363 |
| 6.0 | 93,940 |
| 6.8 | 91,738 |
| 7.1 | 98,273 |
| 7.9 | 101,302 |
| 8.2 | 113,812 |
| 8.7 | 109,431 |
| 9.0 | 105,582 |
| 9.5 | 116,969 |
| 9.6 | 112,635 |
| 10.3 | 122,391 |
| 10.5 | 121,872 |

Tabel di atas terdiri dari 2 kolom, yaitu ‘Tahun_bekerja’ dan ‘Gaji’, di mana data diurutkan dari tahun bekerja kecil ke besar. Dapat dilihat bahwa semakin lama seseorang bekerja kecenderungannya semakin tinggi pula gajinya. Namun terkadang tahun bekerja yang lama tidak selalu bergaji lebih besar dari pegawai yang bekerja lebih singkat. Misal kita amati ada pekerja yang bekerja selama 3.9 tahun bergaji 63K euro, sementara di atasnya pekerja bekerja selama 4.5 tahun bergaji 61K euro.

Pertanyaan yang muncul, apakah memang hubungan antara lama bekerja dengan besarnya gaji adalah linear? Jika linear, seberapa kuat kelinearitasannya?

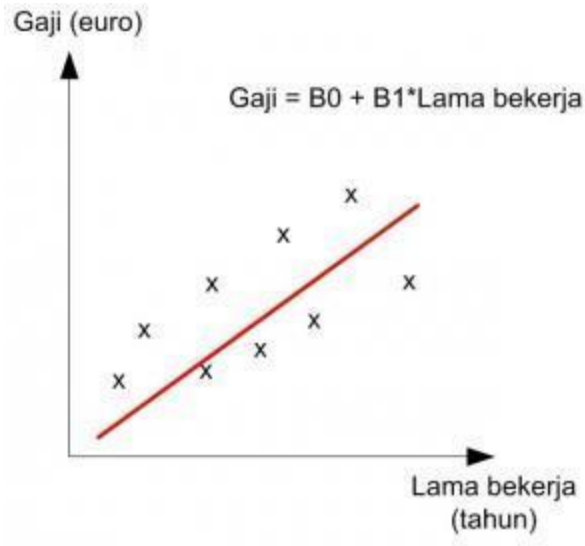
Regresi melalui salah satu tekniknya yaitu s*imple linear regression* (SLR) menjawab pertanyaan di atas. SLR mencari hubungan antara 1 variabel independen (lama bekerja) dengan 1 variabel dependen. Jika variabel independennya lebih dari satu, maka namanya menjadi *multiple linear regression*.

Formula dari SLR diberikan sebagai berikut:




Y adalah variabel dependen, dan X adalah variabel independen. B0 adalah intercept (konstanta), dan B1 adalah slope (koefisien pengali), sementara epsilon adalah error dari sebuah model regresi.

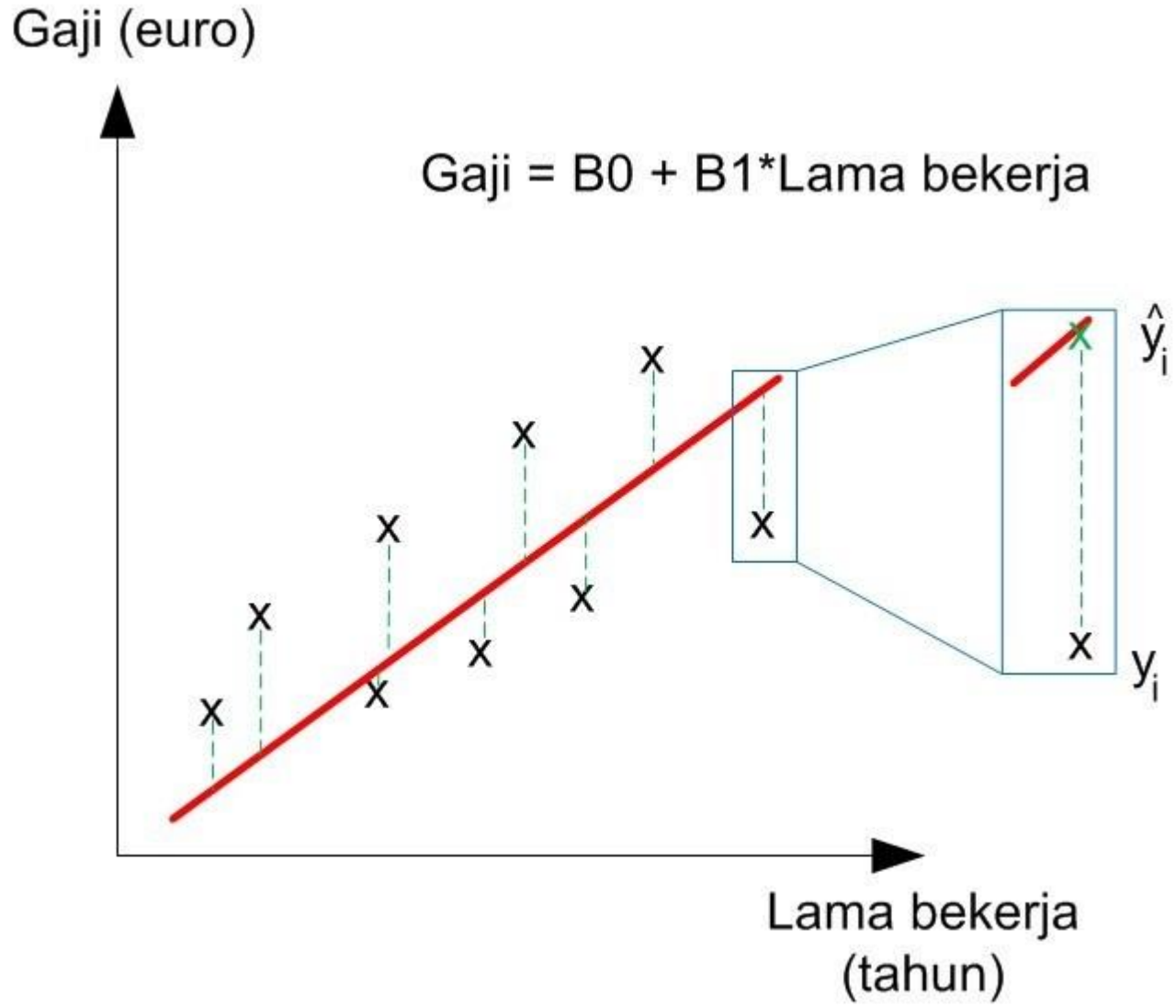
Hasil dari model regresi yang kita buat, kurang lebih akan tampak seperti berikut:




Garis merah merupakan model regresi yang terbentuk. Di mana ia menunjukkan tren positif yang berarti naik ke atas ditunjukkan dengan *slope* (nilai B1 positif). Jika sebaliknya (tren negatif) maka ia cenderung menurun ke bawah (B1 negatif).

Lalu bagaimana sebuah model regresi terbentuk? Metode yang sering digunakan adalah *Ordinary Least Square,* di mana sebuah model akan membuat sebanyak mungkin

garis linear kemudian menghitung selisih kuadrat antara data sesungguhnya terhadap data prediksi model. Ilustrasinya sebagai berikut:




Garis merah merupakan model regresi yang terbentuk. Sementara data real yang kita miliki adalah x, dan kita memiliki x sebanyak 9 buah. Data sesungguhnya ini kita notasikan sebagai Yi, dan data prediksi adalah Yi topi. Selisih keduanya ditandai dengan garis putus-putus warna hijau. Oleh karena itu metode *ordinary least square* didefinisikan sebagai berikut:

$$min \sum (Yi - \widehat{Yi})^2$$

Model regresi akan membuat garis linear sebanyak mungkin, kemudian dari semua garis tersebut dihitunglah nilai *sum squares* nya (formula di atas). Model yang memiliki nilai *sum squares* terkecil merupakan model regresi terbaik yang dipilih.