Analisis Regresi

Disertai Contoh-Contoh Menggunakan R

Tedy Herlambang

2022-01-20

Contents

Kata Pengantar

Manusia sering menghubung-hubungkan suatu hal dengan hal yang lain. Sebaiknya kita menyimpulkan hubungan ini atas dasar bukti-bukti sahih, bukan spekulasi. Salah satu alat yang bisa dipakai untuk membuktikan adanya hubungan antar variabel adalah analisis regresi.

Analisis regresi mengekplorasi hubungan antara satu variabel, biasa disebut sebagai variabel respons dengan satu atau lebih variabel prediktor/penjelas

kemudian mengekspresikan hubungan ini dalam sebuah fungsi. Pengetahuan dan ekspresi hubungan ini dapat dipakai untuk membuat prediksi dan mengidentifikasi variabel-variabel yang paling berpengaruh terhadap sebuah variabel respons. Sebagai misal kita memprediksi keberhasilan studi dan prestasi kerja seseorang (respons) berdasarkan IQ, jenis kelamin, latar belakang sosial-ekonomi dan pendidikan ortunya (prediktor). Prediksi menguji kemampuan kita memetakan variabel-variabel yang paling berpengaruh terhadap suatu variabel respons. Jika kita tidak tepat mengidentifikasi variabel-variabel ini, prediksi dan skenario yang kita buat bisa meleset.

Analisis regresi mencakup metode grafik dan analitik. Dengan bantuan komputer dan pemikiran yang cermat, analisis regresi dapat menjadi alat untuk membedah data yang kompleks menjadi informasi yang berguna. Tetapi, komputer tidak bisa berpikir dan merencanakan sendiri. Komputer hanya membantu kita melakukan perhitungan dengan cepat, agar kita memiliki lebih banyak waktu untuk berpikir dan memperbaiki model analisis.

Di dalam buku singkat ini, saya akan membahas analisis regresi dan penerapannya. Pembahasan dimulai dengan regresi linier sederhana, dilanjutkan ke regresi berganda, permasalahan yang biasa kita temui dalam analisis regresi serta analisis regresi lanjutan dan ditutup dengan diskusi. Program yang akan digunakan adalah R (**?**) yang bisa diunduh secara gratis di sini.

Tetapi buku ini bukan buku tentang R. Juga bukan buku tentang statistik. Materi untuk belajar R sangat banyak tersedia secara *online* misalnya Introduction to R. Pembaca juga bisa mengikuti berbagai topik di R-bloggers. Tujuan buku ini adalah sebagai referensi untuk memperkuat pemahaman pembaca tentang model regresi sebagai salah satu alat analisis yang bisa diterapkan pada berbagai bidang akademik maupun praktik. Saya mengharapkan buku ini bisa menjadi *live book* yang berkembang sesuai saran pembaca. Buku ini adalah versi *online* dari versi cetak yang akan diterbitkan kemudian.

Cara mengutip:

> Herlambang, Tedy. 2022. Analisis Regresi. URL https://bangtedy.github.io/analisisregresi.

Disclaimer

Terakhir diperbarui pada: 20 January 2022.

Chapter 1

Pendahuluan

Sebagai praktisi atau akademisi, terkadang kita tertarik untuk meningkatkan pemahaman atas sesuatu dengan menginvestigasinya. Tertulis atau hanya tercatat di dalam pikiran, di dalam proses investigasi ini, kita menyusun hipotesis-hipotesis yang ingin kita buktikan kebenarannya.

Untuk sampai pada kesimpulan menolak atau menerima hipotesis-hipotesis ini, kita merancang investigasinya, mengumpulkan data/bukti-bukti kemudian menganalisisnya. Di dalam proses analisis ini, seorang investigator biasanya memiliki data sampel yang berasal dari suatu populasi. Hasil analisisnya dapat menyediakan bukti-bukti statistik yang mendukung atau menolak hipotesis yang telah dibuat serta menampilkan karakteristik-karakteristik penting dari populasi dimana data yang dianalisis berasal.

Proses investigasi ini biasanya iteratif, tidak sekali luncur langsung selesai. Dalam beberapa hal, investigasi harus kita ulang lagi sampai didapatkan sebuah model statistik yang memuaskan dengan mempertimbangkan kepraktisan, biaya dan waktu yang tersedia.

Subjek investigasi atau percobaan kita dapat berupa hal-hal yang berada di dalam atau di luar diri kita, yang kasatmata atau tidak. Aerobiologi mempelajari partikel-partikel organik (bakteri, spora jamur, polen, serangga-serangga kecil) yang terbawa udara secara pasif. Astronomi mempelajari benda langit dan fenomena-fenomena alam yang terjadi di luar atmosfer bumi. Ilmu Ekonomi mempelajari kegiatan produksi, alokasi dan distribusi barang dan jasa. Psikologi mempelajari alam pikiran manusia. Geologi mempelajari bentuk dan komposisi bumi.

Analisis regresi tidak mempelajari suatu subjek secara langsung tetapi sebagai alat analisis yang dipakai di sebuah bidang ilmu. Kegunaannya tidak langsung tetapi melalui bantuan yang diberikan ke bidang ilmu lain.

Analisis regresi sangat berguna, sebab hampir semua cabang ilmu pengetahuan

harus berhubungan dengan data yang tidak sempurna. Ketidaksempurnaan data ini mungkin terjadi karena kita hanya dapat mengamati dan mencatat sebagian saja dari apa yang relevan dengan subjek penelitian kita.

Atau bisa juga karena kita hanya dapat mengamati secara tidak langsung dari apa yang benar-benar relevan dengan subjek yang diamati. Mungkin juga karena seberapapun hati-hati kita melakukan observasi atau mendesain sebuah percobaan, data yang diperoleh akan selalu mengandung unsur 'gangguan' (*noise*).

Analisis regresi adalah salah satu teknik statistik yang paling luas dipakai (**?**), (**?**) dan (**?**). Penerapannya meliputi hampir semua bidang ilmu: bisnis, ekonomi, teknik, ilmu-ilmu sosial, biologi dan kesehatan. Pada beberapa proyek penelitian analisis regresi bahkan seringkali menjadi alat analisis utamanya. Keberhasilan penerapan model regresi linier memerlukan pemahaman baik akan teori yang mendasarinya maupun persoalan-persoalan praktis yang sering terjadi di dalam penggunaan alat analisis ini dalam situasi riil.

1.1 Model Regresi Linier

Model di dalam analisis regresi merujuk pada ekspresi matematik yang menjelaskan perilaku-perilaku dari variabel-variabel yang menjadi perhatian. Model regresi linier atau biasa disebut analisis regresi (**?**) digunakan untuk menjelaskan atau memodelkan hubungan antara satu variabel respons y dengan satu atau lebih variabel prediktor $x_1, x_2, ..., x_p$, dimana p adalah jumlah prediktor. Secara khusus, analisis regresi adalah upaya untuk menjelaskan pergerakan sebuah variabel respons dengan merujuk pada pergerakan satu atau lebih variabel eksplanatori. Jika $p = 1$, modelnya disebut regresi sederhana. Tetapi jika $p > 1$, disebut regresi berganda.

Dalam analisis regresi, kita dapat menggunakan pengetahuan tentang hubungan ini untuk memprediksi respons variabel y melalui variabel x. Karena hubungan inilah maka y disebut variabel respons dan x disebut variabel prediktor. Rumpun ilmu berbeda kadang menggunakan istilah yang berbeda untuk menyebut variabel x dan y, seperti yang ditunjukkan pada Tabel berikut.

Di dalam ekonometrika seringkali digunakan istilah variabel dependen dan variabel independen. Di bidang ilmu yang berhubungan dengan eksperimen biasanya digunakan istilah variabel respons dan variabel kontrol karena variabel x berada dibawah kendali peneliti. Dibuku ini penyebutan-penyebutan itu akan digunakan secara bebas.

Jika memungkinkan, di dalam analisis regresi kita juga ingin menyimpulkan apakah terdapat hubungan sebab-akibat: mengetahui pengaruh dari variabel prediktor terhadap variabel respons.

Beberapa contoh kasus analisis regresi misalnya:

Variabel y	Variabel x
Dependen	Independen
Endogen	Eksogen
Respons	Perlakuan/Kontrol
Regressand	Regressor
Ruas kiri persamaan	Ruas kanan persamaan
Diprediksi	Prediktor
Output	Input
Dijelaskan	Penjelas/Eksplanatori

1. Hubungan antara ukuran kelas/jumlah siswa per kelas dengan rata-rata nilai pelajaran Bahasa Daerah di sebuah SMA. Disini y adalah rata-rata nilai pelajaran Bahasa Daerah dan x adalah jumlah siswa per kelas.
2. Hubungan antara tingkat pendapatan seseorang (y) dengan pendidikan terakhir yang ditamatkannya (x).
3. Kinerja karyawan dapat diprediksi dengan menggunakan hubungan antara kinerja (y) dan hasil tes aptitutenya (x).
4. Apakah terdapat hubungan antara tingkat kelahiran penduduk suatu negara dengan tingkat pendapatan perkapitanya? Jika diperhatikan negara-negara dengan tingkat pendapatan per kapita tinggi (x), tingkat pertumbuhan penduduknya rendah (y), dan sebaliknya.
5. Jumlah kosakata seorang anak dapat diprediksi dengan menggunakan pengetahuan hubungan antara jumlah kosa kata (y), usia anak (x_1) dan tingkat pendidikan orang tua si anak (x_2).
6. Hubungan antara desain pekerjaan (x) dengan perilaku karyawan di sebuah perusahaan (y). Perusahaan menginginkan agar karyawan beraktivitas dan berinteraksi secara efektif sehingga pekerjaan harus didesain dengan baik sehingga menghasilkan perilaku positif dan komitmen tinggi karyawan terhadap pekerjaannya. Walaupun inti investigasi ini adalah analisis regresi, tetapi di dalam contoh ini, kita berhubungan dengan konsep-konsep abstrak/konstruk-konstruk yang memerlukan alat analisis sendiri seperti analisis faktor. Topik ini dibahas di buku saya yang lain tentang analisis faktor.

Sebagai ilustrasi, kita akan menggunakan kasus hubungan antara tingkat penjualan sepeda motor dengan pertumbuhan pendapatan per kapita di Indonesia dalam 20 tahun terakhir. Motor sebagai salah satu moda transportasi sangat populer di Indonesia karena berbagai alasan: praktis dan terjangkau. Penjualan motor berhubungan dengan tingkat pendapatan, harga motor, ketersediaan moda transportasi alternatif, selera, dll. Di kasus ini kita hanya mengambil pendapatan sebagai variabel eksplanatori.

Tabel berikut menunjukkan data dalam kasus ini. Data penjualan motor (y) diperoleh dari AISI, sedangkan data pendapatan didekati dengan pertumbuhan pendapatan per kapita (x) yang diperoleh dari World Bank. Satuan y dalam

Tahun	y	x	Tahun(lanj.)	y(lanj.)	x(lanj.)
2001	1.575822	2.235180	2011	8.012540	4.748318
2002	2.287706	3.090636	2012	7.064457	4.606485
2003	2.809896	3.376533	2013	7.743879	4.151428
2004	3.887678	3.630909	2014	7.867195	3.639072
2005	5.074186	4.289592	2015	6.480155	3.555063
2006	4.428274	4.107514	2016	5.931285	3.758837
2007	4.688263	4.946468	2017	5.886103	3.841197
2008	6.215830	4.620034	2018	6.383108	3.987825
2009	5.951962	3.247328	2019	6.487460	3.871444
2010	7.369249	4.812273	NA	NA	NA

juta unit, sedangkan x dalam persen. Secara teoritis terdapat hubungan positif antara tingkat pendapatan dengan penjualan motor karena motor adalah barang normal.

Langkah awal untuk melihat apakah memang terdapat "sinyal" yang menunjukkan hubungan antara pendapatan dengan penjualan motor dari data yang kita miliki adalah dengan mengamati data secara berpasangan dan membuat diagram pencarnya (*scatter plot/scatter diagram*). Setiap data y dipasangkan dan diplot terhadap nilai x-nya.

Diagram pencar bisa menunjukkan secara kasar apakah hubungan antar variabel dapat dianggap linier atau tidak. Jika nilai y cenderung meningkat atau menurun secara garis lurus ketika nilai x meningkat, dan jika perpencaran pasangan titik-titik (x, y) berada disekitar garis lurus, maka kita dapat menjelaskan hubungan antara y dan x dengan menggunakan model regresi linier. Hasil plot pada Gambar 1.1 menunjukkan bahwa memang terdapat sinyal yang sesuai dengan asumsi yaitu penjualan motor proporsional dengan pertumbuhan pendapatan per kapita: semakin tinggi pertumbuhan pendapatan per kapita semakin tinggi pula tingkat penjualan motor.

Selanjutnya kita tambahkan sebuah garis lurus yang paling banyak "mendekati" titik-titik data pengamatan yang tersebar. Sebuah garis lurus yang kita tarik seperti pada Gambar 1.2 menunjukkan bahwa proporsionalitas ini memang ada tetapi tidak ketat. Garis lurus yang kita tambahkan tidak menyinggung seluruh data yang tersebar. Dengan kata lain data sampel tidak bisa seluruhnya tepat berada pada garis yang dibuat. Titik-titik pengamatan ada yang tepat digaris, ada juga yang di atas atau di bawah garis.

Ini menunjukkan adanya **variasi** pada penjualan motor yang tidak berhubungan dengan tingkat pendapatan. Atau terdapat faktor-faktor lain selain pendapatan yang juga mempengaruhi tingkat penjualan tetapi tidak tercakup di dalam model.

Jika kita ekspresikan ke dalam persamaan, garis regresi yang memodelkan hubungan antara y dan x dapat kita tuliskan sebagai:

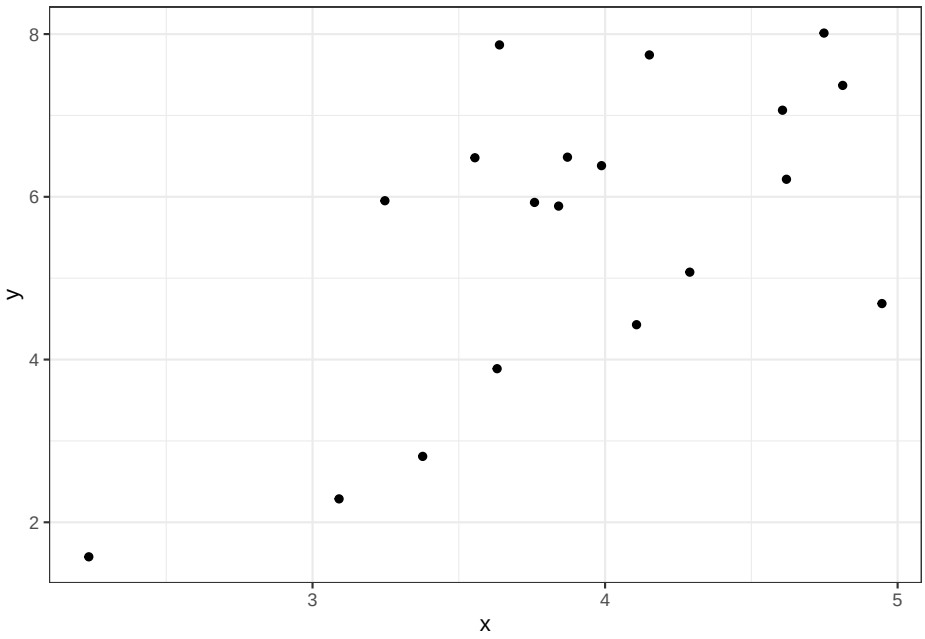

Figure 1.1: Diagram Pencar Penjualan Motor (y) sebagai Fungsi dari Pendapatan per Kapita (x)

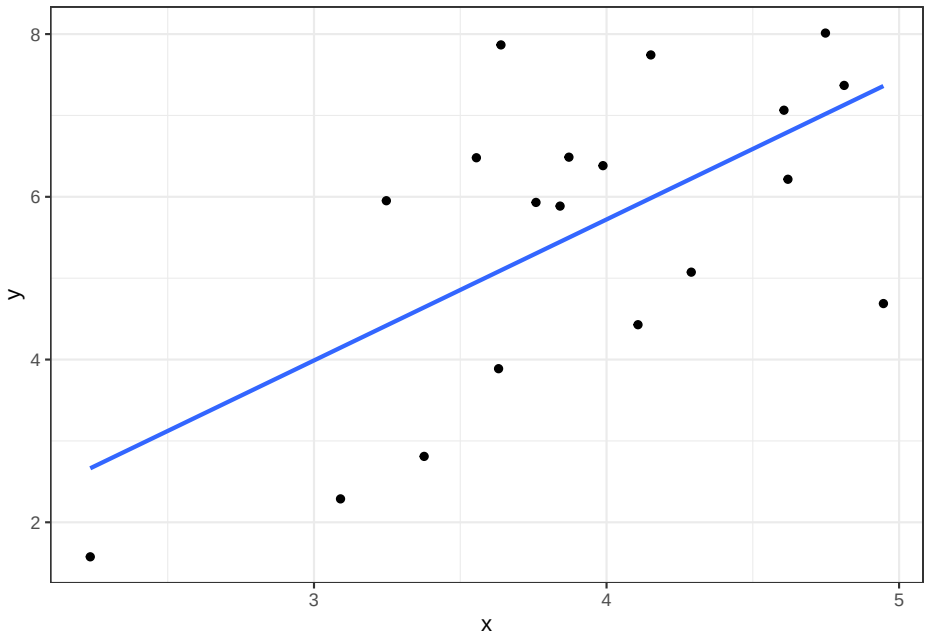

Figure 1.2: Garis Regresi Penjualan Motor (y) sebagai Fungsi dari Pendapatan per Kapita (x)

$$y_i = \beta_0 + \beta_1 x_i \tag{1.1}$$

Persamaan ini menyatakan bahwa y sebagai variabel respons adalah fungsi linier dari x, variabel prediktor. β_0 adalah intersep (*intercept*) yang menunjukkan nilai dari y jika x sama dengan nol. β_1 adalah koefisien kemiringan garis (*slope*) yang menunjukkan berapa banyak y akan berubah jika x meningkat sebanyak satu unit.

Di dalam analisis regresi kita mencari nilai dugaan $\hat{\beta}_0$ dan $\hat{\beta}_1$ sehingga rata-rata jarak vertikal untuk setiap titik data pengamatan dengan nilai dugaannya secara kolektif paling kecil (Gambar 1.3). Dengan kata lain kita ingin menarik garis regresi dalam bidang x−y sedekat mungkin dengan persebaran titik-titik data sampel yang telah kita kumpulkan sehingga variasi yang tidak dapat dijelaskan oleh model menjadi minimum.

Salah satu cara untuk mendapatkan garis yang paling pas (*line of best fit*) yang sangat populer adalah melalui suatu prosedur formal yang disebut *metode kuadrat terkecil/ordinary least squares* (OLS). Sesuai namanya metode ini meminimalkan jumlah kuadrat kesalahan dari model.

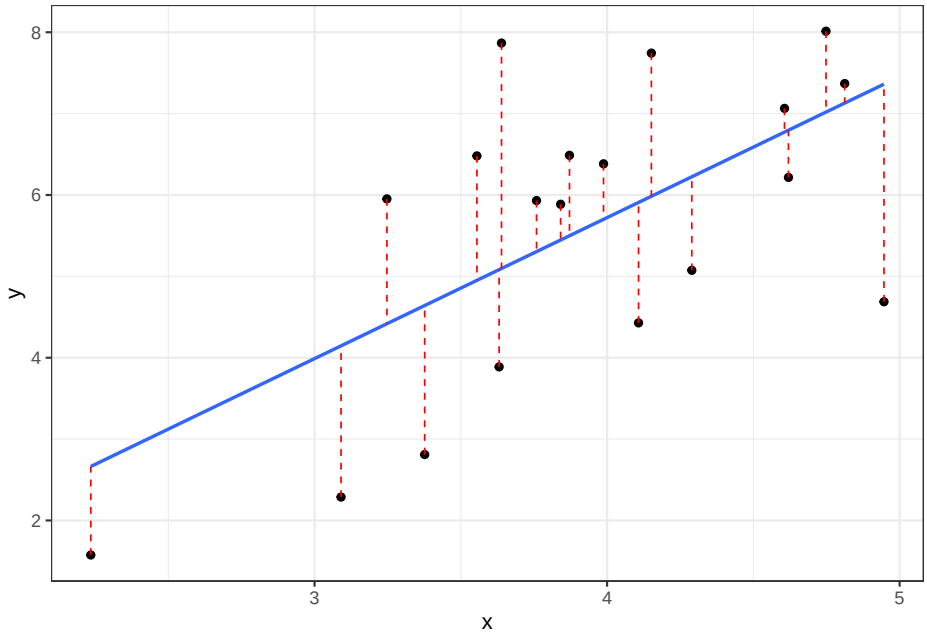

Figure 1.3: Mencari Garis Regresi dengan Metode Kuadrat Terkecil

Di dalam kasus penjualan motor, karena kita hanya menggunakan satu variabel prediktor dalam regresi sederhana, faktor lain selain pendapatan yang menye-

babkan variasi pada penjualan motor kita masukkan sebagai faktor acak (*random factor*). Faktor acak ini secara visual ditunjukkan oleh panjang garis vertikal (warna merah putus-putus) antara titik data pengamatan dengan garis regresi atau nilai dugaan variabel respons (garis biru).

Faktor acak ini menunjukkan adanya variasi pada penjualan yang tidak dapat dijelaskan oleh model dan disimbolkan dengan ϵ. Maka persamaan garis regresi dapat kita tuliskan lagi sebagai:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{1.2}$$

ϵ_i dimasukkan ke dalam persamaan untuk mengakomodasi variasi dalam y yang tidak dapat dijelaskan oleh variabel x. Faktor acak ini adalah istilah statistik yang mewakili fluktuasi acak, kesalahan pengukuran, atau dampak dari faktor-faktor diluar kendali peneliti (**?**) yang menyebabkan nilai dugaan tidak sama dengan nilai aktual (*error* atau residu). Jadi ϵ ini mewakili ketidakmampuan model untuk menjelaskan variasi yang terjadi pada variabel respons.

Setiap model pasti mengandung *error*, karena hubungan statistik tidak memiliki ketepatan seperti halnya fungsi matematika. Jadi ketika kita mencoba menjelaskan variabel y dengan menggunakan variabel x ini, kita menggunakan hubungan statistik: sebuah hubungan yang tidak eksak.

Bahkan, seandainya variabel-variabel lain ditambahkan ke dalam model, tetap akan ada variasi di dalam y yang tidak dapat dijelaskan oleh model. Variasi ini bisa karena bentuk fungsional yang tidak benar atau bisa juga karena semata-mata faktor kejadian yang tidak dapat diprediksi.

Sebuah model regresi linier yang sudah dikalibrasi paling pas dengan data sampel (*the best fit*), dapat digunakan untuk menjelaskan hubungan atau membuat prediksi dengan mencermati perbedaan (arah) dan intensitas perubahan (magnitud) dari variabel y pada nilai variabel x tertentu.

Analisis regresi menawarkan pendekatan yang masuk akal dengan mengenali pola-pola hubungan antar variabel: arah dan besar perubahan pada variabel respons dengan arah dan besar perubahan pada variabel prediktor dari model yang valid.

Dengan demikian persamaan regresi yang memodelkan hubungan statistik antar variabel terdiri dari dua komponen: komponen deterministik atau sinyal yaitu $\beta_0 + \beta_1$ dan komponen acak yaitu ϵ. Kombinasi kedua komponen ini menjadikan persamaan regresi sebagai model non-eksak, yaitu bergantung pada bentuk-bentuk penjelasan probabilistik. Kita hanya mengasumsikan bahwa di dalam cakupan variabel yang dianalisis, terdapat tendensi variabel respons y untuk bervariasi secara sistematis dengan variabel prediktor x.

Mayoritas analisis regresi yang baku menggunakan model regresi linier, yaitu mengasumsikan bahwa variabel respons dapat dituliskan sebagai kombinasi linier dari variabel-variabel prediktornya. Beberapa alasan mengapa model linier

ini paling umum dipakai adalah (**?**): (1) model linier mudah dipahami; (2) beberapa model nonlinier secara instrinsik linier, sehingga bisa didekati dengan pendekatan linier.

Dalam banyak hal, jumlah variabel respons yang kita analisis bisa lebih dari satu atau $p > 1$, sehingga persamaan umum regresinya menjadi:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i \tag{1.3}$$

Model ini disebut sebagai model regresi linier berganda. Disebut berganda karena melibatkan lebih dari satu variabel prediktor. Contoh-contoh penerapannya:

1. Seorang mahasiswa ingin menduga koefisien-koefisien dari sebuah model yang menghubungkan bobot tanaman vaskular-tanaman yang memiliki jaringan khusus yaitu xilem dan floem untuk mengangkut unsur hara, air dan mineral ke seluruh bagian tanaman-dengan kandungan unsur hara dalam tanah, jumlah air yang diterima dan jangka waktu tanaman terpapar sinar matahari.
2. Kerja fisik (mengangkat, memutar, mendorong benda berat) tetap tidak dapat dihindari walaupun kita sudah mampu membuat alat-alat untuk membantu pekerjaan. Kadang pekerjaan itu harus kita lakukan dalam kondisi yang tidak ergonomis. Cedera tulang punggung seringkali ditemui pada pekerja di sektor ini. Seorang manajer produksi mungkin tertarik untuk mengurangi masalah cedera tulang belakang ini dengan menginvestigasi hubungan kepadatan mineral tulang belakang dengan usia, berat dan tinggi badan, jenis kelamin dan gaya hidup pekerja.
3. Seorang analis ingin mengetahui tingkat kepuasan karyawan sebuah perusahaan terhadap pekerjaannya. Skor total kepuasan kerja karyawan dihitung dari penjumlahan skor-skor 20 pertanyaan dengan menggunakan 5 poin skala likert. Sedangkan prediktornya digunakan variabel demografi meliputi usia, jenis kelamin dan pendidikan terakhir karyawan.
4. Di dalam teori konsumen, preferensi seorang konsumen terhadap suatu produk dipengaruhi oleh pengetahuan konsumen terhadap produk, emosi,*word of mouth*, faktor-faktor personal dan lingkungan.

1.2 Asumsi-Asumsi Model Linier

All models are wrong, but some are useful (**?**). Seperti halnya metodologi statistik yang lain, analisis regresi linier dapat menjadi cara yang sangat efektif untuk memodelkan data sepanjang asumsi-asumsinya terpenuhi. Jika asumsinya tidak terpenuhi, metode kuadrat terkecil berpotensi memberikan hasil yang arahnya tidak tepat (*misleading*).

Setelah analisis regresi dilakukan, kita harus melakukan uji diagnostik: memastikan apakah model memenuhi asumsi-asumsi model linier. Uji diagnostik dapat dilakukan secara grafik maupun dengan uji formal.

Asumsi-asumsi regresi linier adalah:

1. Linieritas data: hubungan antara prediktor x dengan respons y diasumsikan linier.
2. Normalitas residu. error/residu diasumsikan terdistribusi secara normal.
3. Homogenitas variansi residu: residu diasumsikan mempunyai variansi yang tetap (homoskedastisitas)
4. Independensi residu.

Potensi-potensi tidak terpenuhinya asumsi model regresi adalah:

1. Hubungan antara respons dengan prediktor tidak linier
2. Heteroskedastisitas: variansi residu tidak tetap.
3. Adanya nilai-nilai yang "berpengaruh" besar yang berasal dari (a) nilai pencilan (*outlier*) yaitu nilai-nilai ekstrim pada variabel respons; (b) high-leverage points: nilai-nilai ekstrim pada variabel prediktor.

Uji hipotesis, interval dan prediksi didasarkan atas kepercayaan bahwa asumsi-asumsi model regresi dipenuhi. Jadi penting sekali untuk melakukan pengecekan terhadap asumsi-asumsi ini. Jika asumsi-asumsi dipenuhi, berbagai bentuk inferensi seperti prediksi, kontrol, ekstraksi informasi, penemuan pengetahuan dan evaluasi risiko dapat dilakukan dengan kerangka argumen deduktif sesuai dengan model yang dibangun.

1.3 Uji Signifikansi

Salah satu fungsi dari regresi linier adalah untuk estimasi kondisi populasi. Kita mengamati dan mengumpulkan data sampel, tetapi kita ingin mengetahui apakah data sampel yang kita miliki juga menerangkan sesuatu tentang populasi dimana data ini diambil. Dengan kata lain kita ingin mengetahui apakah hasil dari analisis data sampel dapat digeneralisasi ke populasi dengan uji signifikansi.

Uji signifikansi dapat dilakukan jika asumsi-asumsi model regresi telah dipenuhi. Uji-uji signifikansi itu antara lain:

1. Uji t: digunakan untuk mengetahui apakah koefisien-koefisien regresi secara statistik berbeda secara signifikan (*statistically-significantly*) dari nol.
2. Uji F: jika uji t digunakan untuk menguji hanya satu koefisien, uji F digunakan untuk menguji lebih dari satu koefisien secara serempak.

1.4 Langkah-langkah Melakukan Analisis Regresi

Berdasarkan uraian sebelumnya, secara umum langkah-langkah dalam melakukan analisis regresi adalah sebagai berikut:

1. Menentukan variabel respons y yang akan kita pelajari atau buat modelnya.
2. Menentukan sejumlah variabel prediktor yang kita anggap berguna di dalam menjelaskan variabel respons.
3. Mengumpulkan data (sampel) yang dapat digunakan untuk menguji model.
4. Mengestimasi model.
5. Cek kecukupan model/uji diagnostik (jika hasilnya kurang memuaskan, kembali ke tahap 1).
6. Uji signifikansi dan inferensial.
7. Menulis hasil analisis.

1.5 Penggunaan Komputer

Jika diperhatikan langkah-langkah untuk melakukan analisis regresi pada 1.3, membangun model regresi merupakan sebuah proses iteratif. Dimulai dengan kajian teori yang berhubungan topik yang sedang diteliti dan ketersediaan data untuk menentukan variabel respons dan variabel eksplanatori untuk membangun model awal.

Salah satu pertimbangan penting di dalam memilih variabel prediktor adalah apakah variabel tersebut dapat mengurangi variasi dalam variabel respons. Pertimbangan lain adalah seberapa mudah, murah dan akurat data variabel itu bisa diperoleh dibandingkan calon variabel prediktor yang lain. Pemilihan ini harus cermat karena bagaimanapun model adalah sebuah penyederhanaan dari realitas yang lebih kompleks, sehingga sebaiknya beberapa prediktor saja dimasukkan ke dalam model.

Menampilkan data dalam grafik atau diagram pencar seringkali sangat berguna untuk spesifikasi model awal. Setelah itu parameter-parameter model diestimasi. Setelah itu kecukupan model dievaluasi yang meliputi mencari kemungkinan terjadi kesalahan spesifikasi model, kemungkinan tidak memasukkan variabel penting atau memasukkan variabel yang tidak perlu, menemukan data pencilan (*outlier*).

Kecukupan dan kecocokan model harus dicek karena menentukan apakah model yang dibuat dapat dipakai atau tidak. Hasil dari cek kecukupan mungkin

mengindikasikan apakah model yang dibuat cukup masuk akal atau perlu dimodifikasi. Di dalam uji kecukupan terutama yang perlu dicek adalah residu sebagai realisasi dari kesalahan model ϵ.

Jika model tidak cukup memenuhi, maka perlu dilakukan tindakan perbaikan dan pendugaan paramater diulang lagi. Proses ini mungkin perlu diulang beberapa kali sampai diperoleh model yang memuaskan. Selanjutnya dilakukan validasi untuk memastikan bahwa model dapat diterima di dalam tahap akhir penerapannya.

Dengan demikian, analisis regresi seringkali melibatkan banyak penghitungan, apalagi jika jumlah sampelnya besar dan variabel prediktornya banyak. Untuk membantu mempercepat proses ini kita menggunakan program komputer.

Program komputer yang bagus adalah alat yang diperlukan dalam proses membangun model. Tetapi program komputer saja belum cukup. Analisis regresi memerlukan seni dan inteligensia dalam penggunaan komputer. Analis harus belajar bagaimana menginterpretasikan output komputer dan mengintegrasikan informasi yang didapat dengan model-model yang akan dibuat selanjutnya.

Berbagai program statistik seperti SPSS, Stata, Minitab, Eviews, SAS, JMP, R, dan lain-lain dapat melakukan penghitungan regresi secara cepat dengan hasil yang kurang lebih sama. Di buku ini saya menggunakan software R (**?**) untuk membuat grafik maupun penghitungan regresinya.

Chapter 2

Regresi Linier Sederhana

Bab ini akan membahas regresi linier sederhana. Istilah regresi sederhana tidak merujuk pada kenaifan penelitiannya tetapi merujuk pada model yang hanya terdiri dari satu variabel respons dan satu variabel prediktor.

Situasi ini sering terjadi pada penelitian sains. Misalnya seorang peneliti ingin memprediksi laju reaksi kimia karena perubahan temperatur, atau ingin mengetahui hubungan antara perubahan diet dengan tingkat kolesterol pada seseorang. Jika dapat diasumsikan bahwa variabel-variabel ini terhubung secara linier, kita dapat menggunakan regresi linier sederhana untuk mengkuantifikasi hubungan ini.

Analisis regresi digunakan ketika solusi eksak tidak tersedia, dalam arti kita tidak akan dapat menemukan nilai tunggal yang dapat mencakup secara lengkap hubungan antara variabel respons dengan prediktornya. Sehingga disini kita mencoba memprediksi setepat mungkin variabel respons atau memprediksi dengan kesalahan terkecil.

Untuk mencapai tujuan ini, kita menganalisis pola-pola variabilitas pada variabel respons dan mencoba melihat apakah variabilitas ini dapat diprediksi dari variabilitas prediktornya.

2.1 Model Regresi Linier Sederhana

Model regresi linier sederhana dapat dituliskan sebagai berikut:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{2.1}$$

Regresi sederhana mengindikasikan hanya ada satu variabel prediktor x untuk menduga variabel respons y. Linier disini diartikan modelnya linier pada pa-

rameternya dalam hal ini β_0 dan β_1. Jadi model $y_i = \beta_0 + \beta_1 {x_i}^2 + \epsilon_i$ adalah linier pada β_0 dan β_1, sementara model $y_i = \beta_0 + e^{\beta_i x_i} + \epsilon_i$ tidak linier.

Misalkan kita memiliki pasangan-pasangan data sampel sebanyak n yang diambil secara acak dari populasi yang lebih besar $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$. Tujuan dari analisis regresi linier adalah menemukan model terbaik yaitu menemukan nilai β_0 and β_1 yang menghasilkan garis paling cocok dengan titik-titik data yang kita punyai.

Dengan kata lain tujuan dari analisis regresi adalah mengestimasi koefisien regresi untuk variabel prediktor sehingga didapatkan nilai dugaan variabel respons sedekat mungkin nilainya dengan nilai pengamatan aktualnya.

Di dalam analisis regresi, model terbaik ditunjukkan oleh garis lurus yang menghubungkan rata-rata variabel prediktor dengan variabel respons sedemikian rupa sehingga jumlah kuadrat kesalahan (jarak vertikal antara titik-titik data pengamatan aktual y_i dengan nilai dugaannya $\hat{y}_i$) minimal.

Untuk memperoleh nilai dugaan β_0 dan β_1 yang paling cocok, kita menggunakan metode kuadrat terkecil (*method of least squares*). Dengan pendekatan kuadrat terkecil, kita mencari nilai dugaan β_0 dan β_1 yang meminimalkan jumlah kuadrat residu/kesalahan $(y_i - \hat{y}_i)$.

2.1.1 Kasus 1: Penjualan Motor dan Pertumbuhan Pendapatan Perkapita

Kita akan melanjutkan kasus hubungan antara penjualan motor dengan pendapatan pada Bab 1 sebelumnya sebagai ilustrasi regresi linier sederhana. Data dapat diunduh di sini. Langkah awal untuk melihat bagaimana hubungan antar variabel adalah membuat diagram pencar.

Plot sangat penting di dalam regresi. Pemeriksaan diagram pencar secara teliti harus mendahului penghitungan regresi. Diagram pencar dapat mengindikasikan apakah model regresi yang diinginkan mungkin masuk akal atau tidak. Kesepakatan dalam membuat diagram pencar, variabel x sebagai variabel penjelas diplot pada sumbu horisontal. Varibel y sebagai variabel respons diplot pada sumbu vertikal.

Untuk membuat diagram pencar, hal pertama yang harus dilakukan adalah memasukkan data ke dalam R, mengecek apakah data yang kita masukkan sudah betul dan memanggil *package* yang relevan dengan model yang akan dibuat.

```
# memanggil data yang disimpan dalam bentuk teks ke dalam R
PJMotor <- read.delim("PJMotor.txt")
PJMotor <- as.data.frame(PJMotor)
```

```
# untuk melihat beberapa baris data teratas
head(PJMotor)
```

```
##   Tahun        y        x
## 1  2001 1.575822 2.235180
## 2  2002 2.287706 3.090636
## 3  2003 2.809896 3.376533
## 4  2004 3.887678 3.630909
## 5  2005 5.074186 4.289591
## 6  2006 4.428274 4.107514
```

```
# untuk melihat beberapa baris data terakhir
tail(PJMotor)
```

```
##    Tahun        y        x
## 14  2014 7.867195 3.639072
## 15  2015 6.480155 3.555062
## 16  2016 5.931285 3.758837
## 17  2017 5.886103 3.841197
## 18  2018 6.383108 3.987825
## 19  2019 6.487460 3.871444
```

Data yang kita masukkan kelihatan seperti yang diharapkan. Variabel penjualan dan pendapatan semua dibaca sebagai angka (*numeric data type*). Selanjutnya kita akan lihat struktur dan ringkasan persebaran datanya.

```
# melihat struktur data
str(PJMotor)
```

```
## 'data.frame':    19 obs. of  3 variables:
##  $ Tahun: int  2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 ...
##  $ y    : num  1.58 2.29 2.81 3.89 5.07 ...
##  $ x    : num  2.24 3.09 3.38 3.63 4.29 ...
```

Hasil fungsi `str()` menunjukkan bahwa data ini adalah data *time series* terdiri dari tiga kolom yaitu penjualan motor (y), pertumbuhan pendapatan per kapita (x) dan tahun selama 19 tahun dari 2001-2019.

```
# melihat ringkasan data kecuali data tahun
summary(subset(PJMotor, select = c(y,x)))
```

```
##        y               x
##  Min.   :1.576   Min.   :2.235
```

```
##  1st Qu.:4.558   1st Qu.:3.593
##  Median :5.952   Median :3.871
##  Mean   :5.587   Mean   :3.922
##  3rd Qu.:6.776   3rd Qu.:4.448
##  Max.   :8.013   Max.   :4.946
```

Selanjutnya kita cek normalitas data, korelasi dan diagram pencar antara pasangan data y dengan x seperti yang ditunjukkan pada Gambar 2.1.

```
library (GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
GGally::ggpairs(subset(PJMotor, select = c(y,x)))
```

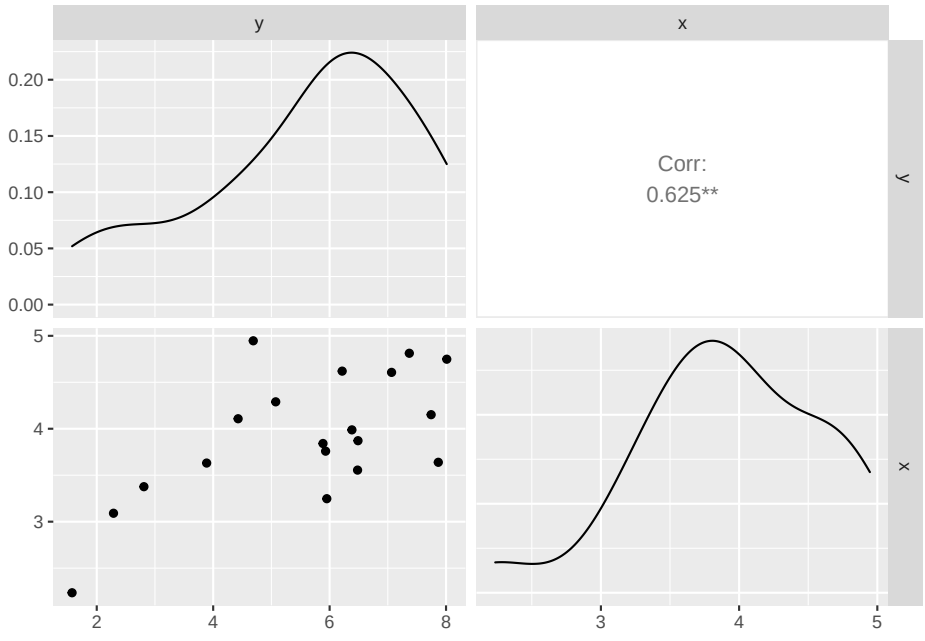

Figure 2.1: Plot Pasangan Data Penjualan dan Pendapatan

Pada diagonal, kita melihat distribusi data y dan x: data kedua variabel menceng ke kiri (*skewed left*). Di atas diagonal ditunjukkan nilai korelasi antara y dan x yaitu sebesar=0.625. Koefisisen korelasi nilainya antara -1 (berkorelasi negatif sempurna) melalui 0 (tidak ada korelasi) sampai $+1$ (berkorelasi positif

sempurna). Korelasi antara penjualan dan pendapatan disini positif cukup kuat dan signifikan. Grafik dibawah diagonal menunjukkan diagram pencar antara pasangan data y dengan x seperti yang sudah kita bahas sebelumnya di Bab 1.

Di dalam R model regresi linier sederhana dapat diperoleh dengan perintah `lm(y~x)`. Tanda ~ dapat diartikan y dijelaskan oleh x atau y fungsi dari x. Fungsi ini mengestimasi koefisien regresi model linier dengan metode kuadrat terkecil (*the least squares method*).

Semisal modelnya kita beri nama **model penjualan motor (mpm)**, yang memodelkan hubungan antara penjualan motor (y) dengan pertumbuhan pendapatan (x), maka model ini dapat diperoleh dengan perintah:

```
mpm <- lm(y ~ x)
# Perintah ini sama dengan y = 0 + 1x
# Intersep secara default diestimasi.
# Bandingkan dengan perintah untuk mendapatkan diagram pencar.
```

Kita kemudian menggunakan perintah `summary()` untuk menampilkan luarannya. Hasilnya adalah sebagai berikut:

```
# perintah summary() untuk mengekstrak hasil regresi
summary (mpm)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6738 -1.1721  0.2916  0.9911  2.7707
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.210      2.088  -0.579  0.56999
## x              1.733      0.525   3.301  0.00422 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.515 on 17 degrees of freedom
## Multiple R-squared:  0.3906, Adjusted R-squared:  0.3547
## F-statistic: 10.89 on 1 and 17 DF,  p-value: 0.004224
```

Persamaan regresinya dapat diringkas sebagai $\hat{penjualan} = -1.21 + 1.73 pendapatan$. Nilai dugaan intersep garis regresi $\hat{\beta}_0$ = -1.210 (SE=

2.088), p-value = 0.57 (tidak signifikan). Nilai -1,21 menunjukkan dugaan penjualan sepeda motor (y dalam juta unit) jika pertumbuhan pendapatan per kapita sebesar 0 persen. Angka penjualan tidak mungkin negatif, paling rendah adalah 0. Jadi dalam kasus ini, nilai intersep dari model yang kita duga tidak mempunyai interpretasi yang berarti. Jadi intersep di dalam sebuah persamaan kadang-kadang mempunyai interpretasi yang berarti kadang juga tidak seperti dalam kasus kita ini.

Sedangkan nilai dugaan variabel prediktor pendapatan $\hat{\beta}_1 = 1.733$ (SE = 0.525) dan p-valuenya (menguji hipotesis nol bahwa $\beta_1 = 0$) kecil (signifikan), konsisten dengan bukti persebaran data yang cenderung linier. Jadi 1.733 adalah nilai kemiringan garis regresi yang menunjukkan untuk setiap perubahan pertumbuhan pendapatan per kapita sebesar 1%, berkaitan dengan perubahan penjualan motor sebesar 1.73 juta unit. Jika terjadi kenaikan pertumbuhan pendapatan per kapita sebesar 1 persen, ini berkaitan dengan peningkatan penjualan motor sebanyak 1.73 juta unit.

Output lain yang penting dari perintah `summary()` adalah koefisisen determinasi (R^2) sebesar 0.39. Ini berarti pendapatan dapat menjelaskan sebesar 39% variasi pada data penjualan motor. Dengan kata lain jika kita ingin menjelaskan mengapa penjualan motor naik turun, maka kita bisa melihat variasi pada pertumbuhan pendapatan. Tentu saja ada faktor lain yang menjelaskan fluktuasi penjualan motor. Tetapi karena model kita hanya memasukkan pendapatan sebagai variabel eksplanatori, maka cukup masuk akal jika model ini hanya menjelaskan sebanyak 39%. Artinya terdapat 61% variasi pada penjualan motor yang tidak bisa dijelaskan oleh pendapatan perkapita saja.

Selang kepercayaan batas atas dan batas bawah (95% defaultnya) untuk nilai dugaan koefisien dapat diperoleh dengan perintah `confint()`.

```
# perintah menampilkan selang kepercayaan
confint(mpm)
```

```
##                 2.5 %   97.5 %
## (Intercept) -5.615467 3.196075
## x            0.625206 2.840601
```

Terlihat bahwa selang kepercayaan untuk intersep meliputi angka 0 (tidak signifikan), sedangkan koefisien untuk pendapatan tidak meliputi 0 (signifikan).

2.1.1.1 Residu

Selisih antara nilai data aktual penjualan motor y_i dengan nilai dugaannya $\hat{y}_i$ pada saat pendapatan $x = x_i$ disebut residu. Nilai residu mencerminkan kegagalan garis regresi yang diestimasi untuk memodelkan pasangan data tersebut. Bagaimana nilai residu diperoleh secara detil?

$\hat{\beta}_0$ dan $\hat{\beta}_1$ adalah nilai-nilai koefisien garis regresi dengan menggunakan data sampel sebanyak 19 tahun $(x_1, y_1), (x_2, y_2), ..., (x_{19}, y_{19})$. Untuk setiap pertumbuhan pendapatan per tahun sebesar x, nilai dugaaan penjualan motornya adalah sebesar $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. Untuk setiap pertumbuhan pendapatan sebesar x_i, dimana $i = 1, 2, ..., 19$, nilai selisih antara $y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ disebut residu.

Nilai residu ini dapat diperoleh dengan menggunakan perintah `augment()` dari *broom package*. Misalkan residu dari perintah `augment()` kita sebut sebagai uji.diagnostik (dibahas secara lebih lengkap di Bab 4):

```
# menambahkan nilai-nilai dugaan dan residu
# library perlu dipanggil jika belum dipanggil sebelumnya
# library("broom")
uji.diagnostik <- augment(mpm)
head(uji.diagnostik)
```

```
## # A tibble: 6 x 8
##       y     x .fitted .resid   .hat .sigma .cooksd .std.resid
##   <dbl> <dbl>   <dbl>  <dbl>  <dbl>  <dbl>   <dbl>      <dbl>
## 1  1.58  2.24    2.66  -1.09 0.394    1.52  0.277      -0.922
## 2  2.29  3.09    4.15  -1.86 0.136    1.48  0.136      -1.32
## 3  2.81  3.38    4.64  -1.83 0.0883   1.49  0.0776     -1.27
## 4  3.89  3.63    5.08  -1.19 0.0628   1.53  0.0222     -0.814
## 5  5.07  4.29    6.22  -1.15 0.0689   1.53  0.0228     -0.786
## 6  4.43  4.11    5.91  -1.48 0.0568   1.51  0.0304     -1.01
```

Kolom-kolom pada tabel diatas menunjukkan

```
x: pertumbuhan pendapatan per kapita
y: penjualan motor aktual
.fitted: nilai dugaan penjualan motor
.resid: nilai residu (penjualan motor aktual-nilai dugaan penjualan motor)
```

Kode berikut memplot nilai residu (warna merah) yaitu selisih antara nilai pengamatan aktual dengan nilai dugaan. Setiap garis vertikal warna merah menunjukkan nilai residu antara data penjualan motor aktual dengan nilai dugaannya.

```
# library tidak perlu dipanggil lagi jika sudah dipanggil sebelumnya
# library(ggplot2)
# library(ggpmisc)
rumus <- y ~ x
ggplot(uji.diagnostik, aes(x, y)) +
  geom_point() +
```

```
  stat_smooth(method = "lm", se = FALSE, formula = rumus, size = 0.8) +
  geom_segment(aes(xend = x, yend = .fitted), color = "red", size = 0.4, linetype = "da
  theme_bw()+
   stat_poly_eq(formula = rumus,
                eq.with.lhs = "italic(hat(y))~`=`~",
                aes(label = paste(..eq.label.., ..rr.label.., sep = "~~~")),
                parse = TRUE)
```

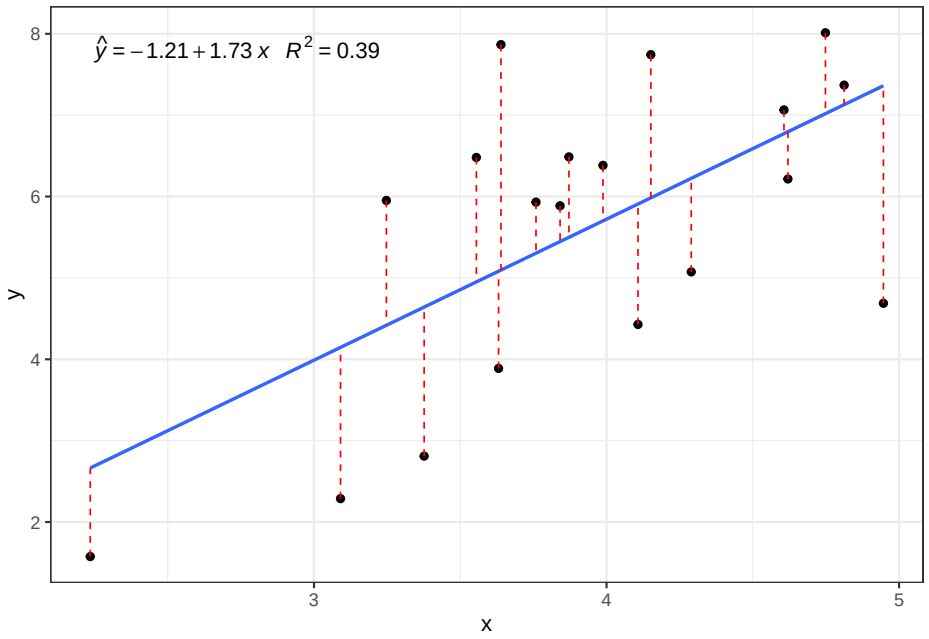

Figure 2.2: Ilustrasi Geometrik dari OLS. Jarak antara Pengamatan dan Garis Regresi Sepanjang Sumbu y

work in progress...sedang dalam proses pengerjaan

Chapter 3

Regresi Linier Berganda

Regresi linier berganda (*multiple linear regression*) adalah pengembangan dari regresi linier sederhana. Dengan regresi linier berganda, kita dapat memasukkan lebih dari satu variabel penjelas ke dalam model. Hal ini dikarenakan di dalam praktik model yang kita pakai kemungkinan melibatkan lebih dari satu variabel prediktor.

Pengembangan ini bermanfaat jika dilihat dari dua sisi. Sisi pertama, penambahan variabel penjelas dapat memberikan penjelasan yang lebih lengkap tentang variabel respons, karena jarang suatu fenomena hanya disebabkan oleh hanya satu hal. Kedua, dampak dari variabel prediktor tertentu dibuat lebih terang, karena kemungkinan dampak distorsi dari variabel prediktor yang lain dihilangkan.

Pemahaman dasar-dasar dari regresi linier sederhana sangat penting untuk memahami regresi linier berganda yang lebih rumit. Dengan bantuan program komputer, proses estimasi dan interpretasi parameter mengikuti prinsip-prinsip yang sama. Begitu juga dengan uji signifikansi, koefisien determinasi (R^2) dan asumsi-asumsi pada regresi linier sederhana terus dibawa ke dalam regresi linier berganda.

Hal-hal yang harus diperhatikan karena berpotensi menjadi masalah dalam melakukan analisis regresi adalah isu-isu yang berhubungan (1) *overfitting*, (2) heteroskedastisitas, (3) multikolinearitas. Teknik regresi berganda sangat luas cakupannya. Penguasaan teknik regresi berganda akan memberikan bekal yang sangat berharga untuk menganalisis berbagai jenis data kuantitatif.

3.1 Persamaan Umum

Secara umum, di dalam persamaan regresi berganda variabel respons dipandang sebagai fungsi linier dari *lebih dari satu* variabel prediktor $1 > p$.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i \tag{3.1}$$

Untuk model dengan dua variabel prediktor saja, persamaannya dapat dituliskan sebagai:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \tag{3.2}$$

yang menunjukkan bahwa y ditentukan oleh x_1 dan x_2, ditambah faktor kesalahan. Untuk mengestimasi nilai-nilai parameter, kita gunakan prinsip kuadrat kesalahan terkecil (*the least squares principle*), dengan meminimalkan jumlah kuadrat kesalahan prediksi (SSE):

$$SSE = \Sigma(y - \hat{y})^2 \tag{3.3}$$

Koefisien-koefisien yang diperoleh yaitu $(\beta_0, \beta_1, \beta_2)$ menghasilkan kesalahan prediksi paling kecil dibandingkan kombinasi-kombinasi koefisien yang lain. Tetapi disini kita tidak dapat lagi menampilkan diagram pencar pada bidang dua dimensi. Kita harus menampilkan perpencaran datanya pada bidang tiga dimensi. Lokasi dari garis pada bidang tiga dimensi ini ditentukan oleh besaran nilai-nilai $(\beta_0, \beta_1, \beta_2)$. Jika variabel penjelasnya lebih dari tiga, perpencaran datanya sangat sulit untuk dibayangkan atau digambarkan.

sedang dalam proses pengerjaan (*work in progress…*)

Chapter 4

Uji Diagnostik dan Tindakan Perbaikan

Di dalam analisis regresi, kriteria jumlah kuadrat terkecil (OLS) tidak akan memberikan hasil yang memuaskan kecuali asumsi-asumsinya dipenuhi. Sampai saat ini kita belum melakukan uji asumsi-asumsi OLS. Uji diagnostik digunakan untuk melihat apakah asumsi-asumsi model dipenuhi atau terjadi pelanggaran pada asumsi-asumsi tersebut. Uji diagnostik biasanya menggunakan nilai residu/error.

Di bab ini kita akan menggunakan alat-alat yang dapat digunakan untuk mengidentifikasi dan mengatasi permasalahan-permasalahan yang biasa ditemui dalam penerapan metode kuadrat terkecil. Kita akan menggunakan baik cara grafik maupun rumus-rumus di dalam uji diagnostik ini.

4.1 Uji Asumsi dengan Plotting Nilai Residu

Hanya dengan melihat beberapa plot nilai residu bukti-bukti pemenuhan/pelanggaran asumsi OLS bisa kita dapatkan. Hal ini menjadi keunggulan uji diagnostik menggunakan grafik yaitu fleksibilitas. Hasil plot dapat menunjukkan bukti-bukti pelanggaran asumsi-asumsi serta tidak memerlukan spesifikasi pasti mengenai bentuk pelanggarannya. Fleksibilitas ini menjadi keunggulan sekaligus menjadi kelemahannya. Cara ini bersifat subjektif, sehingga orang berbeda bisa mempunyai pandangan yang berbeda pula mengenai validitas asumsi-asumsinya. Selain itu plot hanya bisa memberikan "pandangan" dua dimensi dari regresi berganda.

Beberapa plot untuk menguji asumsi OLS:

1. Plot residu dengan nilai dugaan.

2. Plot residu dengan masing-masing prediktor.
3. Plot residu dengan waktu untuk data yang mengandung struktur waktu.
4. Plot normal residu.

4.1.1 Plot residu dengan nilai dugaan.

4.1.2 Plot residu dengan masing-masing prediktor.

4.1.3 Plot residu dengan waktu untuk data yang mengandung struktur waktu.

4.1.4 Plot normal residu.

4.2 Uji Asumsi dengan Tes Statistik

sedang dalam proses pengerjaan (*work in progress...*)

Chapter 5

Model Regresi Linier Lanjutan

Ketika terjadi perubahan kuantitas sebuah variabel yang ada di dalam sebuah sistem, kita biasanya ingin tahu dampak dari perubahan itu terhadap kuantitas variabel-variabel lain yang ada di sistem tersebut. Hubungan perubahan ini mungkin cukup dimodelkan dengan hubungan fungsional sederhana. Bisa jadi juga hubungan perubahan ini lebih kompleks sehingga hubungan fungsional sederhana harus dikembangkan agar kita dapat menganalisisnya. Persamaan umum model regresi dengan variabel respons y dan prediktor x sebanyak p: $x_1, x_2, ..., x_p$ adalah:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon \tag{5.1}$$

dimana ϵ terdistribusi secara normal. Persamaan di atas dapat dikembangkan berdasarkan tiga bagian penyusunnya. Bagian pertama adalah variabel responsnya y, kedua adalah bagian residunya ϵ, dan ketiga adalah variabel prediktornya x:

1. *Generalized Linear Models (GLM)*: Model linier standard tidak dapat mengakomodasi variabel respons y yang tidak normal seperti data dalam bentuk proporsi, persentase, binari, kategori dan *count*. Di dalam kasus variabel respons yang kita hadapi seperti ini kita gunakan *generalized linear model*.

2. *Mixed Effect Model*: Model ini kita gunakan jika data kita tersusun secara hierarkis atau berkelompok. Misal untuk data pengamatan berulang, panel data, longitudinal dan data berjenjang yang mengakibatkan suatu struktur korelasi pada komponen errornya ϵ.

3. Model Regresi Nonparametrik (*Nonparametric Regression Model*): Pada model linier, variabel prediktor x dikombinasikan secara linier untuk memodelkan dampaknya pada variabel respons. Akan tetapi kadang kala, linieritas ini tidak cukup menangkap struktur data sehingga diperlukan fleksibilitas lebih. Metode-metode yang dapat mengakomodasi ini misalnya *additive model, trees and neural networks* memungkinkan pemodelan yang lebih fleksibel pada respons yang mengkombinasikan prediktor secara *nonparametrik*.

5.1 *Generalized Linear Models (GLM)*

5.2 *Mixed Effect Model*

5.3 Model Regresi Nonparametrik

sedang dalam proses pengerjaan (*work in progress...*)