# KASUS REGRESI: IKUT-IKUTAN MENGHAKIMI ASUMSI DAN MEMPERTANYAKAN UJI SIGNIFIKANSI?

***Soetarlinah Sukadji***

Tidak semua masyarakat penggemar penelitian kuantitatif, memiliki pengetahuan sempurna mengenai matematika yang mendasari statistika. Banyak peneliti, termasuk penulis sendiri, adalah orang-orang yang tidak canggung menggunakan statistika meski hanya sebagai *a true believer*, yaitu penganut setia pendapat pakar-pakar statistika. Walau begitu, kita dapat menggunakan kemampuan *reading comprehension* dan logika kita, untuk merunut pikiran pakar-pakar tersebut: apakah benar logikanya, dan apakah terbukti pada hasil olahan data observasi atau simulasi.

Saya termasuk penganut setia Kang Nurija J. Norusis (/SPSS Inc). Ia telah memperkaya kehidupan saya (melalui projek-projek penelitian dan kesempatan menggurui orang lain), dan memberi fasilitas (SPSS) sehingga kehidupan saya dalam dunia penelitian dan bimbingan skripsi-tesis-disertasi menjadi nyaman. Oleh karena itu pengalaman intim dengan Kang Nurija saya manfaatkan dalam membeberkan kasus ini.

## KENALKAN. INI DIA: REGRESI MAJEMUK

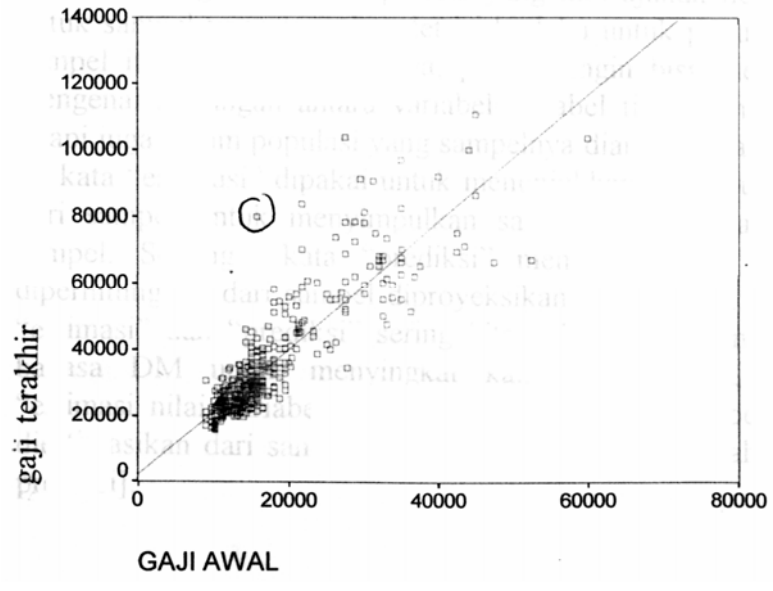
Aliasnya adalah *Multiple Regression,* yang nge-*top* di kalangan mahasiswa. Ada pepatah "tidak dikenal, maka tidak dicinta". Tetapi, pepatah ini tidak berlaku untuk Regresi Majemuk (RM). Banyak mahasiswa tidak kenal, tetapi toh tetap bergelut, bercinta dan memuja-muja RM. Tidak hanya mahasiswa psikologi yang cantik-cantik itu yang gandrung sama RM, tetapi juga mahasiswa ekonomi yang cakep dan merasa sok cakep itu pun, dari S1 sampai S3, jatuh cinta sama RM. Memang RM itu *outlier!* Apa sih rahasianya? Wah, itu sih, tidak semudah diceritakan seperti iklan shampoo. *Multiple regression?* Siapa takut? Hiih.

Apa sih, artinya regresi? Mau tahu saja. Lihat kamus, dong. Tetapi, kira-kira artinya adalah "kembali ke yang sederhana, yang simpel, yang umum". Lha, data yang

dianggap simple dan yang umum, itu yang seperti apa? Gampang jawabnya: yang ke arah *mean*. Ada istilah *regression toward mean.* Contohnya, orangtua yang jangkung-jangkung maupun cebol-cebol, anak-anaknya tidak cenderung makin jangkung atau makin cebol, tetapi justru makin ke arah rata-rata. Itu yang disebut *regression toward mean*. Tentu saja ini tidak dapat diterapkan untuk kekayaan keluarga. *Mean* memiliki *meaning* yang penting. Meramalkan sesuatu menjadi mudah bila kita dapat meramalkan *mean*-nya. Apakah data dari segala macam tingkat pengukuran dapat diramal *mean*-nya? Tentu dong, kecuali bila data itu tipe kategorial. Jadi regresi bisa dipakai untuk data dengan tingkat pengukuran mulai yang ordinal (peringkat) sampai yang rasio. Hebat bukan? Apa pula *outlier?*

## *OUTLIERS* PATUT DICURIGAI, MESKIPUN BUKAN PENIPU

Sebelum kita mengolah, lebih dulu kita cermati datanya, antara lain untuk menemukan *outliers*. Menggunakan *plot* data kita dapat menemukan titik-titik yang letaknya terpencil dari gerombolan data yang lain. Titik-titik yang menyendiri ini patut kita curigai, sebab menyimpang dari kumpulan titik-titik yang lain. Titik-titik yang menyimpang letaknya inilah yang disebut *outliers*. Mengapa patut kita curigai? *Outliers* mungkin terjadi karena kekeliruan pada saat mengumpulkan, koding, atau entre data. Contoh di *plot* Gambar 1 adalah plot titik-titik gaji awal dengan gaji akhir.




**Gambar 1.** Titik terpencil yang disebut *outlier*.

Kita lihat titik yang kita lingkari. Meskipun besarnya gaji awal (Rp. 15.000) dan gaji terakhir (Rp. 80.000) tidak unik, artinya gaji awal maupun gaji terakhir seperti itu umum, tetapi bila keduanya digabung, posisinya menjadi lain daripada yang lain. Ini merupakan salah satu *outliers* yang patut dicurigai.

*Outliers* mungkin sulit diatasi. Seandainya *outliers* tersebut hanya akibat kesalahan koding atau entre data, dengan mudah kita koreksi dan kita *run* kembali analisisnya. Tetapi bila tidak jelas alasannya, kita terpaksa berusaha melihat interaksinya dengan variabel-variabel lain. Misalnya, bila *outlier* mungkin menggambarkan seorang karyawan yang tidak keberatan pada awalnya dipekerjakan sebagai juru tulis bergaji rendah, asal di samping kerja ia dapat meneruskan kuliah MM. Nah, lulus MM, posisinya lalu melejit. Dengan begitu pula dapat kita simpulkan bahwa variabel pendidikan merupakan variabel yang menjelaskan karakteristik gaji karyawan yang unik tersebut. Subjek semacam ini lebih baik tidak kita ikutkan dalam analisis, sebab bisa-bisa *outlier* ini mendominasi hasil analisis.

Mari kita kenali asumsi-asumsi yang mendasari RM melalui bahasan yang disajikan oleh Kang Marija Norusis. Sebetulnya banyak bahasan versi lain, tetapi rasanya edisi yang menjelaskan SPSS 6.1 yang paling enak dilahap. Katanya sih, uji hipotesis dengan regresi mejemuk baru sah bila asumsi-asumsinya dipenuhi.

Kita ingat, bahaw hipotesis yang kita ajukan tidak hanya berlaku untuk sampel yang terjaring, tetapi berlaku untuk populasi yang diwakili sampel itu. Begitulah biasanya, peneliti ingin bisa menarik kesimpulan mengenai hubungan antara variabel-variabel tidak hanya dalam sampel tetapi juga dalam populasi yang sampelnya diambil. (Dalam naskah ulasan ini kata “estimasi” dipakai untuk menunjukkan kesimpulan yang dihitung dari sampel untuk menyimpulkan sampel, yang juga disebut statistik sampel. Sedang kata “prediksi” sering kita pakai tidak mengikuti kaidah tata bahasa DM untuk menyingkat kalimat. Misalnya, yang dimaksud “estimasi nilai variabel dependen” adalah nilai variabel dependen yang diestimasikan dari sampel; “prediksi nilai *mean”* adalah nilai *mean* hasil prediksi).

Untuk membangun garis regresi yang cocok menyimpulkan hubungan dua variabel, selain tingkat pengukuran variabel minimal ordinal, juga perlu diamati apakah hubungan antara kedua variabel linear. Untuk menarik kesimpulan mengenai nilai populasi berdasar hasil sampel, tambahan asumsi berikut ini diperlukan.

- **Semua pengamatan harus independensi.** Artinya, diikutsertakannya satu kasus dalam sampel tidak ada pengaruh keikutsertaan sampel lain. Misalnya, karena seorang anak menjadi sampel, maka kakaknya juga harus menjadi sampel, maka

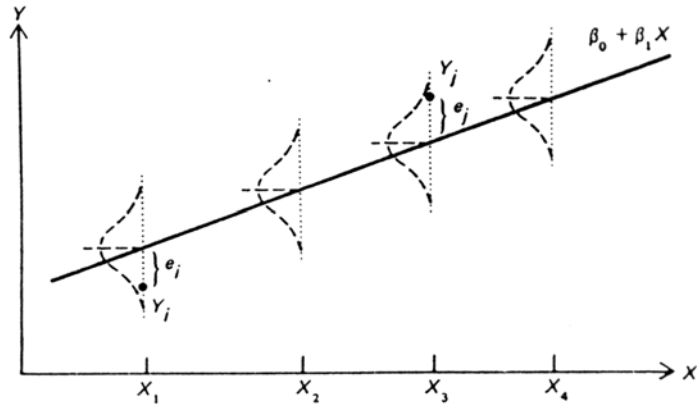
pengamatan ini tidak independen. Jadi. setiap pengamatan sama sekali tidak dipengaruhi oleh pengamatan lain. Contoh lain pengamatan yang tidak independen adalah bila observasi didasari oleh pngukuran berulang (*repeated measurements*) dari unit eksperimen yang sama. Bila dari empat anggota keluarga masing-masing didapat tiga observasi, kedua belas observasi yang diperoleh tidak independen.

- **Untuk setiap nilai variabel independen, distribusi nilai variabel dependen harus normal.** Untuk setiap nilai tertentu variabel independen *X,* distribusi variabel dependen *Y* adalah normal. Kalau kita memperediksikan gaji akhir, asumsi ini menjelaskan bahwa tidak semua karyawan dengan jagi awal sama mendapat gaji akhir yang sama. Gaji akhir untuk setiap gaji awal, bervariasi dengan bentuk distribusi normal.
- **Varians distriusi variabel dependen harus sama untuk setiap nilai variabel independen.** Artinya, distribusi-distribusi normal dependen variabel tersebut di atas memiliki varians yang sama untuk setiap nilai variabel independen. Dengan demikian meskipun *mean* distribusi berbeda-beda, distribusi-distribusi ini memiliki varians yang setara yaitu $\sigma^2$. Makanya disebut juga sebagai "homogenitas varians".
- **Hubungan antara variabel dependen dengan independen dalam populasi harus linear**. Dengan kata lain, *mean* distribusi-distribusi variabel dependen harus terletak pada garis lurus. Ini jelas, karena yang kita hitung adalah regresi linear.

Asumsi-asumsi ini lebih terlihat jelas terlihat pada Gambar 2, yang menggambarkan variabel independennya hanya satu. Model persamaannya:

$$Y_1 = \beta_0 + \beta_1 X_1 + e_1$$

*Slope* dan *intercept* parameter populasi ditunjukkan dengan $\beta_1$ dan $\beta_0$. Nilai $e_i$, yang sering disebut sebagai *error* atau goyangan, adalah perbedaan antara lain pengamatan $Y_i$ dengan *mean* subpopulasi $Y_i$ pada titik $X_i$. Nilai $e_i$ diasumsikan sebagai variabel yang distribusinya normal, independen dan random, dengan *mean* = 0 dan varians $\sigma^2$. Konsep $e_i$ ini penting untuk difahami, sebab akan muncul berulang kali dalam pembahasan regresi. Titik $Y_i$ yang tepat ada di garis regresi, *error*-nya tentusaja sama dengan 0 (ada di *mean* variasi *error*). Makin jauh dari garis regresi makin besar *error*-nya. Mudah-mudahan gambar ini dapat membantu memahami, mengapa asumsi-asumsi ini diperlukan untuk regresi linear.

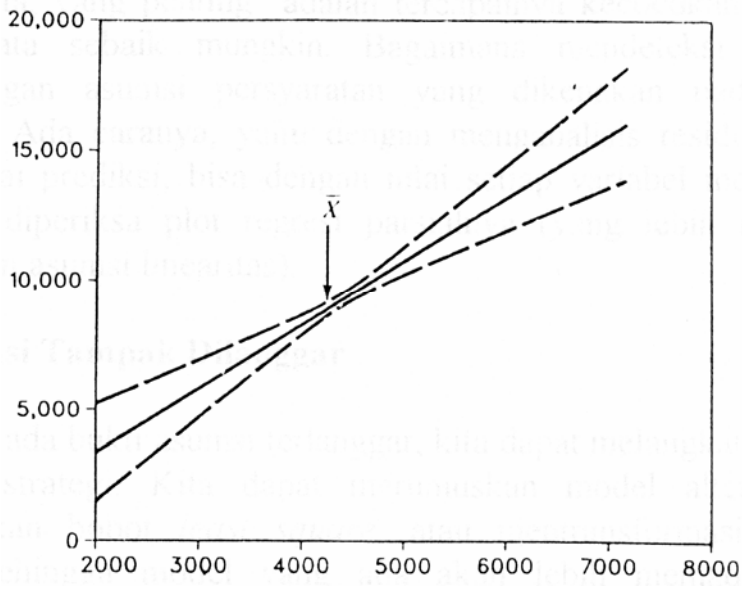



**Gambar 2.** Asumsi Regresi Linear

## CONFIDENCE INTERVAL

Statistik yang dihitung dari sampel menghasilkan suatu estimasi titik yang merupakan estimasi parameter yang tidak diketahui. Suatu estimasi titik ini dapat dipikirkan sebagai tebakan tunggal terbaik untuk nilai populasi. Meskipun nilai hasil estimasi yang diperoleh melalui sampel umumnya berbeda dari nilai parameter populasi yang tidak diketahui, yang diharapkan adalah bahwa nilai tersebut tidak terlalu jauh bedanya. Berdasarkan estimasi sampel tersebut dimungkinkan menghitung rentang nilai yang mencakup nilai populasi tersebut, dengan peluang ketepatannya ditentukan. Rentang semacam itu disebut *confidence interval* (interval kepercayaan). Misalnya, kita ambil contoh Gambar 3, ditentukan *confidence interval* sebesar 95% untuk *mean X* (populasi). Angka rentang yang tertera di situ rentangnya makin menyempit bila mendekati nilai *mean X*, dan makin melebar bila jarak dengan *mean X* bertambah.

*Confidence interval* 95% berarti bahwa bila dari suatu populasi dalam kondisi yang sama, berulangkali diambil sampel, dan 95% *confidence interval* dihitung, 95% interval-interval tersebut akan berisi parameter yang tidak diketahui. Karena nilai parameter tidak diketahui, tidak mungkin dipastikan apakah parameter tersebut tercakup atau tidak dalam interval tertentu.




**Gambar 3.** *Confidence Interval Mean X*.

**Bandingan dengan arti taraf signifikansi dalam uji hipotesis nol.** Misalkan dalam hipotesis nol perbedaan mean ditolak dengan $p \leq 0{,}05$. Interpretasinya di sini lebih langsung, yaitu bahwa probabilitas perbedaan *mean* yang minimal sebesar perbedaan pengamatan itu padahal kedua *mean* sama, hanya mungkin terjadi paling banyak 0,05 kali pengamatan.

Mau menolak taraf signifaknsi? Dari perbandingan kedua paragraf tersebut kita dapat menyimpulkan, kapan digunakan *confidence interval* dan kapan digunakan taraf signifikansi.

## *GOODNESS OF FIT*

Prosedur statistik apa pun yang dipakai untuk menyusun model analisa data, yang penting adalah tercapainya kecocokan antara model dengan data sebaik mungkin. Bagaimana mendeteksi kemungkinan penyimpangan asumsi persyaratan yang dikenakan pada data yang dianalisis? Ada caranya, yaitu dengan menganalisis residu, bisa di-*plot* dengan nilai prediksi, bisa dengan nilai setiap variabel independen, dan bisa pula diperiksa plot regresi parsialnya (yang lebih menampakkan pelanggaran asumsi linearitas).

## BILA ASUMSI TAMPAK DILANGGAR

Bila ada bukti asumsi terlanggar, kita dapat melangkah ke salah satu dari dua strategi. Kita dapat merumuskan model alternatif, seperti menggunakan bobot *least square*, atau mentransformasikan variabel-variabel sehingga model yang ada akan lebih memadai. Misalnya, menggunakan log, akar, atau kebalikan, dapat menstabilkan varians dan mendapatkan normalitas, atau linearitas hubungan.

## MENGUBAH HUBUNGAN NONLINEAR KE LINEARITAS

Dalam usaha mencapai linearitas, kita dapat mentransformasi variabel-variabel independen maupun dependen, atau keduanya. Bila kita mengubah skala variabel-variabel independen, linearitas dapat dicapai tanpa mempengaruhi distribusi variabel dependen. Jadi, bila variabel dependen distribusinya normal dengan varians konstan untuk masing-masing nilai *X*, kondisinya tetap begitu.

Bila variabel-variabel yang kita transformasikan, distribusinya berubah. Distribusi baru ini harus memuaskan asumsi analisis. Misalnya bila log nilai variabel dependen diambil, log *Y*, bukan *Y*, harus normal distribusinya dengan varians konstan.

Pilihan transformasi tergantung pada beberapa pertimbangan. Bila bentuk model-sesungguhnya yang mengarahkan hubungan telah diketahui, kondisi ini mendikte pilihan. Misalnya, bila diketahui bahwa $Y=AC^X$ adalah model yang memadai, menggunakan log untuk kedua sisi persamaan menghasilkan persamaan berikut ini.

$$\log Y_i = \underset{[B_0]}{(\log A)} + \underset{[B_1]}{(\log C)}\, X_i$$

Jadi, hubungan log *Y* dengan *X* linear.

Bila model-sebenarnya tidak diketahui, kita harus memilih transformasi dengan memeriksa plot data. Sering kali, hubungan tampak seperti linear untuk sebagian data tetapi melengkung untuk sisanya. Transformasi lain mungkin mengurangi lengkungan adalah akar *Y,* atau –1/*Y*. Pilihan tergantung pada konteks tertentu, pada beratnya masalah.

## MENGATASI KEJULINGAN

Bila distribusi residu variabel dependen juling kekanan (*positively sweed*, atau kempes sebelah kanan), transformasi log sering dapat mengatasinya. Untuk distribusi yang juling ke kiri, kita biasa memakai transformasinya. Untuk distribusi yang juling ke kiri, kita biasa memakai transformasi kuadrat. Harus diperhatikan bahwa uji-*F*

digunakan dalam uji hipotesis regresi biasanya sangat tidak sensitif terhadap penyimpangan normalitas yang moderat.

## STABILITAS VARIANS

Bila varians residu tidak konstan, kita dapat mencoba berbagai sarana remedial.

- Bila varians residu proporsional dengan *mean Y* untuk *X* tertentu, gunakan akar *Y* bila semua $Y_i$ (prediksi individual) positif.
- Bila deviasi standar residu proporsional dengan *mean*-nya, kita coba menggunakan transformasi logaritma.
- Bila deviasi standar residu proporsional dengan kuadrat *mean*-nya, gunakan kebalikan *Y*.
- Bila *Y* berbentuk proporsi atau peringkat, transformasi ke *arc sinus* mungkin dapat menstabilisasi varians.

## KOMENTAR AKHIR MENGENAI ASUMSI

Jarang ada prosedur analisis regresi atau analisis statistik lain yang tidak melanggar asumsi. Meskipun demikian, ini bukan alasan pembenar untuk mengabaikan asumsi-asumsi. Langsung tancap regresi tanpa memikirkan kemungkinan penyimpangan asumsi-asumsi yang dibutuhkan dapat menyesatkan interpretasi dan pemanfaatan hasil. Taraf signifikansi, *confidence interval*, dan hasil-hasil lain itu peka terhadap tipe pelanggaran tertentu, dan tidak dapat diinterpretasi dengan cara yang sama, bila penyimpangan serius.

Dengan memperhatikan baik-baik residu, dan bila dibutuhkan menggunakan transformasi atau metode analisis yang lain, kita berada dalam posisi yang lebih baik untuk melakukan analisis yang menyelesaikan masalah yang kita teliti. Meskipun tidak semuanya sempurna, kita paling tidak dapat menilai dengan pengetahuan yang kita miliki seberapa kemungkinan kesulitan yang kita hadapi. Amien.

## DAFTAR PUSTAKA

Norusis, M.J. (1990) *SPSS/PC+Statistics™ 4.0.*Chicago: SPSS Inc.

Norusis, M.J. (nd) *SPSS 6.1 Guide to Data Analysis*. Englewood Cliffs, NJ: Prentice Hall.