



MODUL ANALISIS DAN TEKNIK PEMODELAN (MIK 623)



Disusun Oleh
Mieke Nurmalasari, M.Si, M.Sc



UNIVERSITAS ESA UNGGUL
2017



KATA PENGANTAR

Alhamdulillah, puji syukur atas rahmat dan kasih sayang Allah SWT sehingga penulis dapat menyelesaikan pembuatan modul Analisis dan Teknik Pemodelan untuk program studi Manajemen Informasi Kesehatan Universitas Esa Unggul. Penulis mengucapkan terima kasih kepada berbagai pihak atas dukungan, bantuan dan semangat yang diberikan kepada penulis.

Jika terdapat kesalahan penulisan dalam modul ini, penulis akan sangat terbuka dalam menerima masukan, saran dan kritik yang membangun untuk perbaikan modul ini di masa yang akan datang. Semoga modul ini dapat membantu mahasiswa dalam proses pembelajaran di kelas. Terima kasih.

Jakarta, 15 Januari 2018

Penulis

DAFTAR ISI

KATA PENGANTAR

Materi 1: Pengenalan Teknik Pemodelan

Materi 2: Analisis Regresi Linear Berganda

Materi 3: Analisis Regresi Polynomial

Materi 4: Model Buiding

Materi 5: Diagnostik Model

Materi 6: Uji Asumsi Analisis Regresi Berganda

Materi 7: Student Project Discussion

Materi 8: Pengenalan Analisis Data Kategorik

Materi 9: Regresi Logistik Biner Berganda

Materi 10: Analisis Regresi Logistik Multinomial

Materi 11: Analisis Deret Waktu (*Time Series Analysis*)

Materi 12: Teknik Peramalan Deret Waktu (Metode Trend dan Moving Average)

Materi 13: Ukuran Akurasi Model Deret Waktu

Materi 14: Final Student Project

Materi 1

Pengenalan Teknik Pemodelan

A. Pendahuluan

Dalam modul ini akan dijelaskan ulang tentang berbagai macam tipe/ jenis data dan variabel. Hal ini perlu dipaparkan ulang karena sebagai salah satu landasan untuk memilih analisis dan teknik pemodelan yang sesuai. Teknik analisa data yang akan dilakukan untuk mencari informasi yang relevan terhadap dugaan (hipotesis) tergantung pada jenis/ tipe data yang kita miliki. Sehingga sebelum mengenal lebih jauh tentang berbagai macam model-model statistika atau analisa data, maka kenali terlebih dahulu tipe/ jenis data.

B. Kompetensi Dasar

Mengetahui tujuan melakukan analisis data dan membuat pemodelannya, selain itu juga mengenal beberapa teknik pemodelan.

C. Kemampuan Akhir yang Diharapkan

Mahasiswa mampu memahami konsep dasar statistika sebagai landasan dalam membuat model. Dengan mengetahui jenis data yang dimiliki kemudian mengetahui variabel apa saja yang menjadi pusat penelitian serta memahami tujuan dari penelitian maka diharapkan mahasiswa dapat memilih teknik pemodelan yang tepat.

D. Kegiatan Belajar 1

Data dan Variabel

1. Uraian dan contoh

Data merupakan catatan kumpulan fakta yang dipercaya kebenarannya. Data merupakan bentuk jamak dari Datum.

Data dapat dibedakan menurut:

i. Sifatnya

a. Data Kualitatif adalah data dalam bentuk kategorial, contoh kepuasan terhadap pelayanan rumah sakit ada yang sangat puas, puas, cukup dan tidak puas.

b. Data Kuantitatif, terbagi menjadi dua:

Data diskrit adalah data dalam bentuk bilangan bulat atau data yang didapat dari hasil perhitungan. Contoh: jumlah penderita epilepsi, jumlah anak dalam keluarga.

Data kontinyu adalah data dalam bentuk desimal dan didapatkan dari hasil pengukuran. Contoh: berat badan, tinggi badan, panjang badan.

ii. Skala Data

a. Nominal

b. Ordinal

c. Interval

d. Rasio

Skala data yang terendah adalah skala data nominal dan skala data tertinggi adalah skala rasio.

iii. Waktunya

a. Data Silang (*Cross Section*)

b. Data Berkala (*Time Series*)

Variabel adalah sesuatu yang menjadi objek pengamatan penelitian, atau merupakan atribut suatu objek yang memiliki nilai kuantitatif maupun kualitatif yang berbeda-beda satu dan lainnya.

Variabel ada tiga jenis yaitu:

i. Variabel bebas (*independent variable*) merupakan variabel yang mempengaruhi satu atau lebih variabel yang lain. Variabel ini diukur, dipilih atau ditentukan oleh peneliti. Variabel bebas sering disebut juga variabel prediktor atau variabel penjelas.

ii. Variabel tidak bebas (*dependent variable*) merupakan variabel yang dipengaruhi oleh satu atau lebih variabel yang lain. Variabel tidak bebas sering disebut juga variabel target atau variabel respon.

iii. Variabel pengganggu (*confounding variable*) merupakan variabel yang mempengaruhi hubungan antara variabel bebas terhadap variabel tidak bebas secara tidak langsung.

Contoh: Penelitian dampak olahraga (prediktor) terhadap berat badan (target), maka variabel lain seperti usia dan pola makan akan juga mempengaruhi.

2. Latihan

1. Jelaskan data menurut tipe/ jenisnya!

2. Jelaskan dan berikan contoh tentang data *cross section* dan data *time series*!

3. Jelaskan apa yang dimaksud variabel bebas, variabel tidak bebas dan variabel pengganggu, berikan contohnya!

3. Tes Formatif

1. Tentukan apakah data dalam kasus di bawah ini termasuk data diskrit atau kontinyu?

- Jumlah kematian penduduk di wilayah DKI Jakarta tahun 2016.
- Kandungan zat besi dalam sampel darah.
- Jangka waktu penyembuhan setelah serangan jantung
- Banyaknya anak-anak yang dirawat karena demam berdarah di rumah sakit A sepanjang tahun 2016.

2. Tentukan manakah yang termasuk variabel bebas, variabel tidak bebas untuk kasus berikut ini:

“Suatu penelitian ingin mengetahui pengaruh kegiatan merokok terhadap munculnya penyakit kanker paru-paru”

Sebutkan juga apa saja yang dapat menjadi *confounding variable* untuk kasus di atas.

4. Umpan Balik dan Tindak Lanjut

Cocokkan jawaban di atas dengan kunci jawaban tes formatif 1 yang ada di bagian akhir modul ini. Ukurlah tingkat penguasaan materi kegiatan belajar 1 dengan rumus sebagai berikut:

Tingkat penguasaan = (jumlah jawaban benar : 7) x 100%

Arti tingkat penguasaan yang diperoleh adalah:

Baik sekali = 90 – 100%

Baik = 80 – 89%

Cukup = 70 -79%

Kurang = 0 – 69%

Bila tingkat penguasaan mencapai 80% ke atas , silahkan melanjutkan ke Kegiatan Belajar 2. Bila tingkat penguasaan masih di bawah 80% harus mengulangi Kegiatan Belajar 1 terutama pada bagian yang belum dikuasai.

E. Kegiatan Belajar 2

Tujuan dan Jenis Teknik Pemodelan

1. Uraian dan Contoh

Teknik pemodelan adalah suatu metode yang digunakan untuk memecahkan suatu masalah dengan membuat rancangan atau model yang sesuai dengan tujuan atau hipotesis si peneliti.

i. Tujuan Analisis dan Teknik Pemodelan

Tujuan melakukan pemodelan adalah memilih model terbaik dari beberapa kemungkinan model yang bisa dipakai untuk menyelesaikan sebuah masalah.

Terdapat empat jenis analisis data yaitu: analisis deskriptif, analisis komparatif, analisis kecenderungan dan analisis hubungan.

ii. Jenis Teknik Pemodelan: Model Linier dan Model Nonlinier

Model Linier adalah model yang hubungan antara variabel tidak bebas dan variabel bebasnya linier dalam parameter.

Penyelesaian suatu kasus atau penelitian sering dilakukan dengan menggunakan pendekatan model linier atau model non linier yang ditransformasi dengan menjadi model linier. Ada beberapa model nonlinier yang bisa ditransformasi menjadi model linier antara lain model eksponensial, parabolik dll. Tetapi ada beberapa kasus yang tidak bisa lagi dilakukan dengan pendekatan model liner, sebagai alternatifnya dapat digunakan model nonlinier untuk menangani kasus tersebut.

Model Nonlinier merupakan bentuk hubungan antara variabel bebas dengan variabel tidak bebas yang tidak linier dalam parameter.

Model nonlinier dapat dibagi menjadi dua yaitu:

1. Model linier intrinsik (*intrinsically linear*) adalah model nonlinier yang dapat ditransformasi menjadi bentuk linier.
2. Model nonlinier intrinsik (*intrinsically nonlinier*) yaitu model yang tidak bisa ditransformasi menjadi bentuk linier (Draper and Smith, 1981).

Dalam analisis hubungan, pemilihan model sangat dipengaruhi oleh jenis variabel responnya (Y). Secara garis besar dapat disajikan dalam tabel berikut:

Variabel Tidak Bebas (Y)	Variabe Bebas (X)	Distribusi	Model
1 kuantitatif variabel	Kuantitatif, kualitatif atau keduanya	Normal	Regresi linier, regresi non linier
1 kualitatif variabel	Kuantitatif, kualitatif atau keduanya	Binomial	Regresi logistik, multinomial regresi logistik

2. Latihan

1. Jelaskan tentang model linier dan beri contohnya!
2. Jelaskan tentang model non linier dan beri contohnya?

F. Kunci Jawaban

1. Tes formatif 1

1.
 - a. Diskrit
 - b. Kontinyu
 - c. Kontinyu
 - d. Diskrit
2.
 - a. Variabel Bebas: merokok
 - b. Variabel Tidak Bebas: kanker paru-paru
 - c. Variabel Pengganggu: umur, lama merokok, olah raga

Materi 2

Analisis Korelasi dan Regresi Berganda

Pendahuluan

Pada mata kuliah sebelumnya telah dijelaskan mengenai hubungan antara dua variabel serta mengetahui pengaruh suatu variabel terhadap variabel lain. Hubungan dan pengaruh dari dua variabel tersebut bermanfaat untuk mengetahui kondisi atau dampak yang terjadi akibat adanya perubahan suatu variabel terhadap variabel lain, sehingga dapat disusun suatu rencana dalam menghadapi dampak tersebut.

Pada kehidupan nyata, banyak kasus yang memerlukan pengetahuan tentang hubungan tersebut tidak hanya dipengaruhi oleh satu faktor saja melainkan lebih. Oleh sebab itu pada modul ini akan dipelajari tentang analisis korelasi berganda dan analisis regresi berganda untuk mengetahui hubungan dan pengaruh dari beberapa variabel bebas (*independent variable*) terhadap variabel terikat (*dependent variable*).

G. Kompetensi Dasar

Mengetahui cara melakukan analisis korelasi ganda dan regresi berganda sehingga dapat membuat model yang sesuai untuk peramalan.

H. Kemampuan Akhir yang Diharapkan

Mahasiswa mampu memahami konsep dasar analisa korelasi dan analisa regresi berganda sebagai landasan dalam membuat model untuk keperluan prediksi atau peramalan. Mahasiswa juga dapat menghitung secara manual koefisien korelasi ganda dan parsial, koefisien regresi inier ganda dan koefisien determinasi. Selain itu mahasiswa juga mampu melakukan pengujian hipotesis korelasi dan regresi ganda serta dapat menganalisa hasil peramalan dan kemudian menginterpretasikannya.

I. Kegiatan Belajar

Analisis Korelasi dan Regresi Berganda

1. Uraian dan contoh

Sebelumnya akan diulas kembali tentang analisis regresi sederhana dengan rumus umum:

$$\hat{Y} = b_0 + b_1 X_1$$

Persamaan ini mempunyai satu variabel bebas (X) dan satu variabel tidak bebas (Y). Pada kenyataannya dalam kehidupan sehari-hari seringkali nilai Y tidak hanya dipengaruhi oleh satu variabel bebas saja. Jika kita memperhatikan pengaruh dari dua atau lebih faktor (variabel bebas) lainnya maka kita gunakan analisis regresi berganda. Sehingga bentuk umum persamaan regresi dengan k-variabel bebas dapat dimodelkan sebagai berikut:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_k X_k$$

Estimasi Koefisien Regresi

Ada beberapa cara pendekatan dalam menghitung koefisien regresi berganda, diantaranya melalui pemecahan persamaan normal dan pendekatan matriks. Pada modul ini akan dibahas satu pendekatan saja yaitu melalui penyelesaian persamaan normal dan untuk model regresi dengan dua variabel bebas. Persamaan model regresi berganda dengan dua variabel bebas adalah sebagai berikut:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

Maka estimasi koefisien regresinya adalah sebagai berikut:

$$b_1 = \frac{(\sum x_1 y)(\sum x_2^2) - (\sum x_2 y)(\sum x_1 x_2)}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2}$$

$$b_2 = \frac{(\sum x_2 y)(\sum x_1^2) - (\sum x_1 y)(\sum x_1 x_2)}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2}$$

$$b_0 = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2$$

Dengan:

$$\sum x_1^2 = \sum X_1^2 - \frac{(\sum X_1)^2}{n}$$

$$\sum x_2^2 = \sum X_2^2 - \frac{(\sum X_2)^2}{n}$$

$$\sum y^2 = \sum Y^2 - \frac{(\sum Y)^2}{n}$$

$$\sum x_1 x_2 = \sum X_1 X_2 - \frac{(\sum X_1)(\sum X_2)}{n}$$

$$\sum x_1 y = \sum X_1 Y - \frac{(\sum X_1)(\sum Y)}{n}$$

$$\sum x_2 y = \sum X_2 Y - \frac{(\sum X_2)(\sum Y)}{n}$$

Pengujian Hipotesis terhadap Koefisien Regresi

Pengujian hipotesis koefisien regresi dilakukan untuk melihat pengaruh dari setiap variabel bebas terhadap variabel tidak bebasnya apakah berpengaruh secara signifikan atau tidak. Berikut ini langkah-langkah dalam melakukan pengujian hipotesis koefisien regresi berganda:

Langkah 1. Rumuskan Hipotesis

$$H_0 : \beta_k = 0 \quad ; \text{dimana } k = 0, 1, 2, \dots, k$$

$$H_1 : \beta_k \neq 0 \quad (\text{Bisa juga } H_1 > 0 \text{ atau } H_1 < 0, \text{ tergantung kasusnya})$$

Langkah 2: Menghitung statistik uji

$$t_{hitung} = \frac{b_k}{sb_k} \quad \text{Dimana: } k = 0, 1, 2, \dots, k$$

Langkah 3: Menentukan nilai kritis dari tabel sesuai dengan tingkat nyata (α)

$$t_{tabel} = t_{\alpha, t_{\alpha/2}; df = n - k} \quad (k = \text{banyaknya variabel bebas})$$

Langkah 4. Menarik kesimpulan

Jika $t_{hitung} > t_{tabel}$, maka keputusannya adalah TOLAK H_0

Jika $t_{hitung} < t_{tabel}$, maka keputusannya adalah TERIMA H_0

Pengujian Hipotesis terhadap Persamaan Model Regresi

Pengujian terhadap model regresi ini merupakan uji koefisien regresi secara simultan atau bersama-sama. Tujuan dari melakukan pengujian model regresi adalah untuk mengetahui kelayakan dari model regresi yang kita buat.

Dalam perhitungannya kita membutuhkan analisis varians atau *Analysis of Variance* (ANOVA).

Langkah 1. Rumuskan hipotesis

$$H_0 : \beta_0 = \beta_1 = \beta_2 = \dots = \beta_k = 0 ; \text{dimana } k = 0, 1, 2, \dots, k$$

$$H_1 : \beta_k \neq 0 \text{ (paling tidak ada satu nilai } \beta_k \text{ yang tidak sama dengan 0)}$$

Dengan kata lain bisa juga hipotesisnya sebagai berikut:

$H_0 : \beta_0 = \beta_1 = \beta_2 = \dots = \beta_k = 0$ (tidak ada pengaruh yang signifikan X_k terhadap Y)

$H_1 : \beta_k \neq 0$ (paling sedikit ada satu variabel X yang mempengaruhi secara signifikan variabel Y)

Langkah 2: Menghitung statistik uji

Tabel ANOVA

Sumber Variasi	df	Jumlah Kuadrat	Mean Square	F_{hitung}
Regresi	$k - 1$	$b_1 \sum x_1 y + b_2 \sum x_2 y$	$\frac{JK_{regresi}}{k - 1}$	$F_{hitung} = \frac{MS_{regresi}}{MS_{residual}}$
Residual	$n - k$	$JKT - JK_{regresi}$	$\frac{JK_{regresi}}{n - k}$	
Total	$n - 1$	$\sum y^2$		

Atau bisa dijabarkan sebagai berikut:

$$JK_{regresi} = \left[b_1 \sum X_1 Y - \frac{(\sum X_1)(\sum Y)}{n} \right] + \left[b_2 \sum X_2 Y - \frac{(\sum X_2)(\sum Y)}{n} \right]$$

$$JK_{Total} = \sum y^2 = \sum Y^2 - \frac{(\sum Y)^2}{n}$$

$$F_{hitung} = \frac{MS_{regresi}}{MS_{residual}}$$

Langkah 3: Menentukan nilai kritis dari tabel sesuai dengan tingkat nyata (α)

$$F_{\alpha}; df = (k - 1); (n - k), \text{dimana } k = \text{banyaknya semua variabel}$$

Langkah 4. Menarik kesimpulan

Jika $t_{hitung} > t_{tabel}$, maka keputusannya adalah TOLAK H_0

Jika $t_{hitung} < t_{tabel}$, maka keputusannya adalah TERIMA H_0

Analisis Koefisien Determinasi dan Korelasi Berganda

Koefisien Determinasi Berganda

Koefisien determinasi (R^2) adalah menjelaskan seberapa besar pengaruh variabel bebas (X) terhadap variabel tidak bebas (Y). Jika dalam persamaan regresi linier berganda, misalkan ada dua variabel bebas X_1 dan X_2 , maka jika ingin diketahui berapa proporsi (prosentase) sumbangan X_1 dan X_2 terhadap variasi atau naik turunnya variabel Y secara bersama-sama, maka inilah yang disebut koefisien determinasi berganda.

$$R_{y.12}^2 = \frac{b_1 \sum x_1 y + b_2 \sum x_2 y}{\sum y^2}$$

Atau

$$R_{y12}^2 = 1 - \frac{JK_{residual}}{JK_{total}}$$

Koefisien Korelasi Berganda

$$r_{y.12} = \sqrt{R_{y.12}^2}$$

Koefisien Korelasi Parsial

- Korelasi antara variabel Y dengan variabel X_1 (X_2 dianggap konstan).
- Korelasi antara variabel Y dengan variabel X_2 (X_1 dianggap konstan).
- Korelasi antara variabel X_1 dengan variabel X_2 (Y dianggap konstan).

Contoh Soal:

Sebuah perusahaan farmasi di Jakarta sedang melakukan penelitian tentang biaya promosi, biaya distribusi dan total penjualan obat pereda rasa sakit yang baru. Menurut tim pemasaran besarnya total penjualan dipengaruhi oleh biaya promosi dan biaya distribusi. Data yang berhasil dikumpulkan dalam 5 bulan terakhir adalah sebagai berikut:

Bulan	Biaya Promosi(X1)	Biaya Distribusi(X2)	Total Penjualan(Y)
1	625	450	1400
2	600	425	1300
3	525	400	1250
4	425	350	980
5	500	300	1000

Berdasarkan data di atas:

- d. Tentukan persamaan regresi linier berganda!
- e. Tentukan koefisien korelasi ganda dan koefisien determinasi ganda dan jelaskan artinya!

Jawab:

a. $\hat{Y} = -87,196 + 1,095X_2 + 1,785X_3$

- b. $R_{y.12} = 0,987$, artinya hubungan antara total penjualan , biaya promosi dan biaya distribusi adalah kuat dan positif (ini adalah nilai koefisien korelasi ganda).

Koefisien determinasi ganda = $R_{y.12}^2 = 97,4\%$, artinya sebesar 97,4% dari variabel total penjualan dapat dijelaskan oleh variabel biaya promosi dan biaya distribusi, sedangkan sisanya sebesar 2,6% dari variabel total penjualan dapat dijelaskan oleh variabel lainnya dengan tingkat keyakinan 95%.

2. Latihan

Survey kepuasan pasien (*patient satisfaction survey*). Seorang pegawai administrasi sebuah rumah sakit ingin melakukan studi tentang hubungan antara kepuasan pasien dilihat dari usia pasien dan tingkat keparahan penyakitnya (*severity of illness*). Pegawai administrasi tersebut memilih secara acak 6 orang pasien untuk dijadikan data sampel.

Patient Satisfaction	Age	Severity Illness
48	50	51
57	36	46
66	40	48
70	41	44
89	28	43
36	49	54
46	42	50
54	45	48
26	52	62
77	29	50

Berdasarkan data di atas:

- Tentukan persamaan regresi linier berganda!
- Tentukan koefisien korelasi ganda dan koefisien determinasi ganda dan jelaskan artinya!
- Jika seorang pasien baru masuk rumah sakit dengan umur 35 tahun dan tingkat keparahan penyakitnya 45, hitunglah prediksi tingkat kepuasan pasien baru tersebut!

3. Tes Formatif

Lakukan lanjutan analisis dengan data yang sama pada latihan, yaitu tentang survei kepuasan pasien. Buatlah tabel ANOVA dan lakukan pengujian model persamaan regresi, atau lakukan pengujian koefisien regresi secara simultan!

MATERI 3

Analisis Regresi Polynomial

Pendahuluan

Pada materi ini diharapkan mahasiswa mampu memahami Analisis Regresi Polinomial. Selain itu mahasiswa dapat menguraikan dan memahami penggunaan regresi polinomial dan cara membentuk modelnya.

Pengertian Analisis Regresi Polynomial

- Polynomial regresi adalah suatu model regresi linier yang dibentuk dengan menjumlahkan pengaruh masing-masing peubah bebas yang dipangkatkan meningkat sampai order ke-N-1.

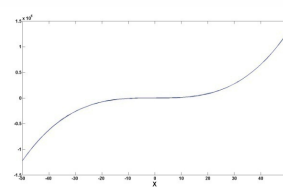
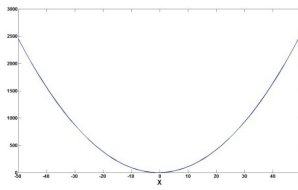
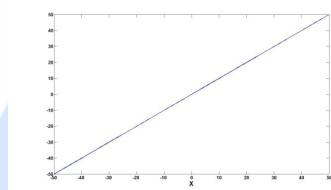
$$y = \beta_0 + \beta_1x + \beta_2x^2 + \beta_3x^3 + \dots + \beta_{N-1}x^{N-1}$$

- PENTING:** Pahami adalah perbedaan antara regressor (parameter) dengan variabel bebasnya.
- Model di atas hanya terdiri dari satu variabel bebas. Tetapi dalam model ini memberi ruang bagi satu variabel bebas (x) ini untuk membentuk hubungan nonlinier antara X dan Y (tetapi modelnya masih merupakan model linier dalam parameternya).

Bentuk Respon Regresi Polynomial

- $Y = \beta_0 + \beta_1x + \beta_2x^2 + \beta_3x^3 + \dots + \beta_{N-1}x^{N-1}$
- Pangkat tertinggi variabel bebas (x) yang menentukan bentuk kurva responnya.

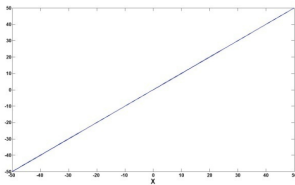
	One bend	Two bends
Linier	Kuadratik	Kubik
$y = \beta_0 + \beta_1x$	$y = \beta_0 + \beta_1x + \beta_2x^2$	$y = \beta_0 + \beta_1x + \beta_2x^2 + \beta_3x^3$
Tidak ada titik belok	Ada satu titik belok	Ada dua titik belok



Tanda dari koefisien regressor dengan pangkat (orde) yang paling tinggilah yang menentukan arah kurva.

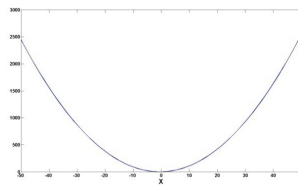
Linier

$$y = 0 + 1x$$



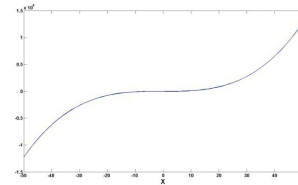
Kuadratik

$$y = 0 + 1x + 1x^2$$

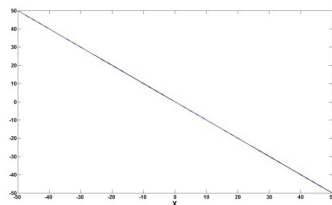


KubikLinear

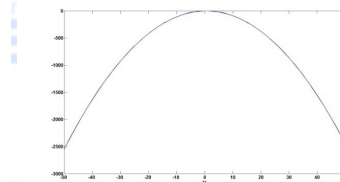
$$y = 0 + 1x + 1x^2 + 1x^3$$



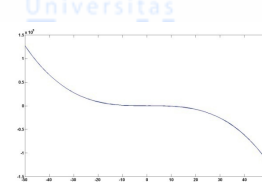
$$y = 0 - 1x$$



$$y = 0 + 1x - 1x^2$$



$$y = 0 + 1x + 1x^2 - 1x^3$$



Kapan kita memakai regresi polynomial?

- Model polynomial sangat berguna ketika si peneliti mengetahui bahwa ada pengaruh kurva linier terhadap responnya bentuknya menyerupai lengkungan.
- Model polynomial juga berguna sebagai fungsi pendekatan untuk model yang sangat kompleks dan hubungannya non linier.
- Model Polynomial merupakan pengembangan seri Taylor (Taylor series expansion) untuk fungsi yang tidak diketahui.

Bentuk Umum Model Regresi Polynomial

- Model regresi linier $y = \beta x + \varepsilon$ adalah model umum dari model linier untuk menjelaskan hubungan linier dari parameter yang tidak diketahui β .

- Polynomial regression model:

- Second order polynomial in one variable

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

- Second order polynomial in two variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon$$

Polynomial Models in One Variable

- A second-order model (quadratic model):

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

- β_1 : Linear effect parameter.
- β_2 : Quadratic effect parameter.

k th order polynomial model in one variable

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \epsilon$$

Contoh Soal

Diberikan data sebagai berikut:

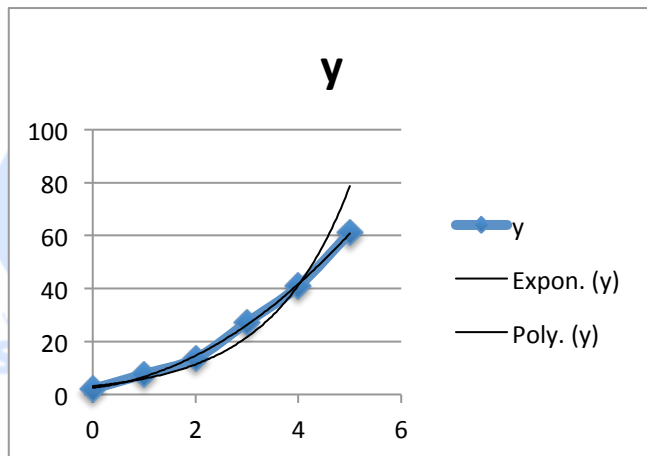
- Buat diagram pencar/ scatter plot dari data !
- Tentukan model regresi yang mendekati sebaran data tersebut!

x	y
0	2.1
1	7.7
2	13.6
3	27.2
4	40.9
5	61.1

Jawaban:

- Diagram pencar/ scatter plot:

Melihat diagram pencar atau scatter plot pada poin a, dapat dilihat bahwa variabel respon (Y) cenderung mengikuti trend kuadratik. Sehingga persamaan regresi yang bisa dibuat untuk mendekati pola di atas adalah regresi polynomial orde-2.



Estimasi koefisien regresi polynomial bisa dilakukan dengan menggunakan pendekatan matrik seperti berikut:

$$\begin{bmatrix} n & \sum X_i & \sum X_i^2 \\ \sum X_i & \sum X_i^2 & \sum X_i^3 \\ \sum X_i^2 & \sum X_i^3 & \sum X_i^4 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \\ \sum X_i^2 Y_i \end{bmatrix}$$

- Masukkan total data di setiap kolom ke dalam matrik:

X	Y	X ²	X ³	X ⁴	XY	X ² Y
0	2.1	0	0	0	0	0
1	7.7	1	1	1	7.7	7.7
2	13.6	4	8	16	27.2	54.4
3	27.2	9	27	81	81.6	244.8
4	40.9	16	64	256	163.6	654.4
5	61.1	25	125	625	305.5	1527.5
ΣX=15 ΣY=152.6 ΣX ² =55 ΣX ³ =225 ΣX ⁴ =979 ΣXY=585.6 ΣX ² Y=2488.8						

$$\begin{bmatrix} 6 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 152.6 \\ 585.6 \\ 2488.8 \end{bmatrix}$$

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 2.47857 \\ 2.35929 \\ 1.86071 \end{bmatrix}$$

Sehingga persamaan regresinya sebagai berikut:

- $y = 2.47857 + 2.35929 x + 1.86071 x^2$

Materi 4

Pembentukan Model Regresi

Pendahuluan

Pada mata kuliah sebelumnya telah dijelaskan bagaimana membuat persamaan regresi berganda dan bagaimana menduga atau mengestimasi nilai koefisien regresinya. Pada bab ini secara garis besar akan dibahas tentang proses pembentukan model dan proses validasi model, kemudian akan dijelaskan lebih detail mengenai beberapa cara dalam pemilihan variabel bebas yang masuk ke dalam model untuk jenis penelitian yang bersifat eksplorasi dan observasional (*exploratory observational studies*). Jika variabel bebas yang masuk ke dalam model banyak maka akan ada banyak kemungkinan model persamaan regresi yang bisa dibentuk. Pada bagian ini akan dijelaskan bagaimana proses dan strategi dalam memilih model terbaik dari beberapa kemungkinan model yang bisa dibentuk. Sehingga akan dipelajari tiga hal mendasar yaitu: prosedur seleksi model, kriteria model dan validasi model.

Kompetensi Dasar

Mengetahui prosedur dalam seleksi model berdasarkan kriteria model yang ada serta menvalidasi model.

Kemampuan Akhir yang Diharapkan

Mahasiswa mampu memahami prosedur dan strategi dalam pemilihan model terbaik dengan melihat kriteria seleksi model.

Kegiatan Belajar

Pendahuluan (Overview of Model Building Process)

Dalam pembuatan model (*model building*) dan pemilihan variabel bebas (prediktor) mana yang masuk ke dalam suatu model memerlukan suatu keahlian khusus dalam pelaksanaannya. Ada beberapa tahapan atau langkah dimana di setiap tahap perlu adanya evaluasi dan pengambilan keputusan yang diperlukan untuk tahap selanjutnya.

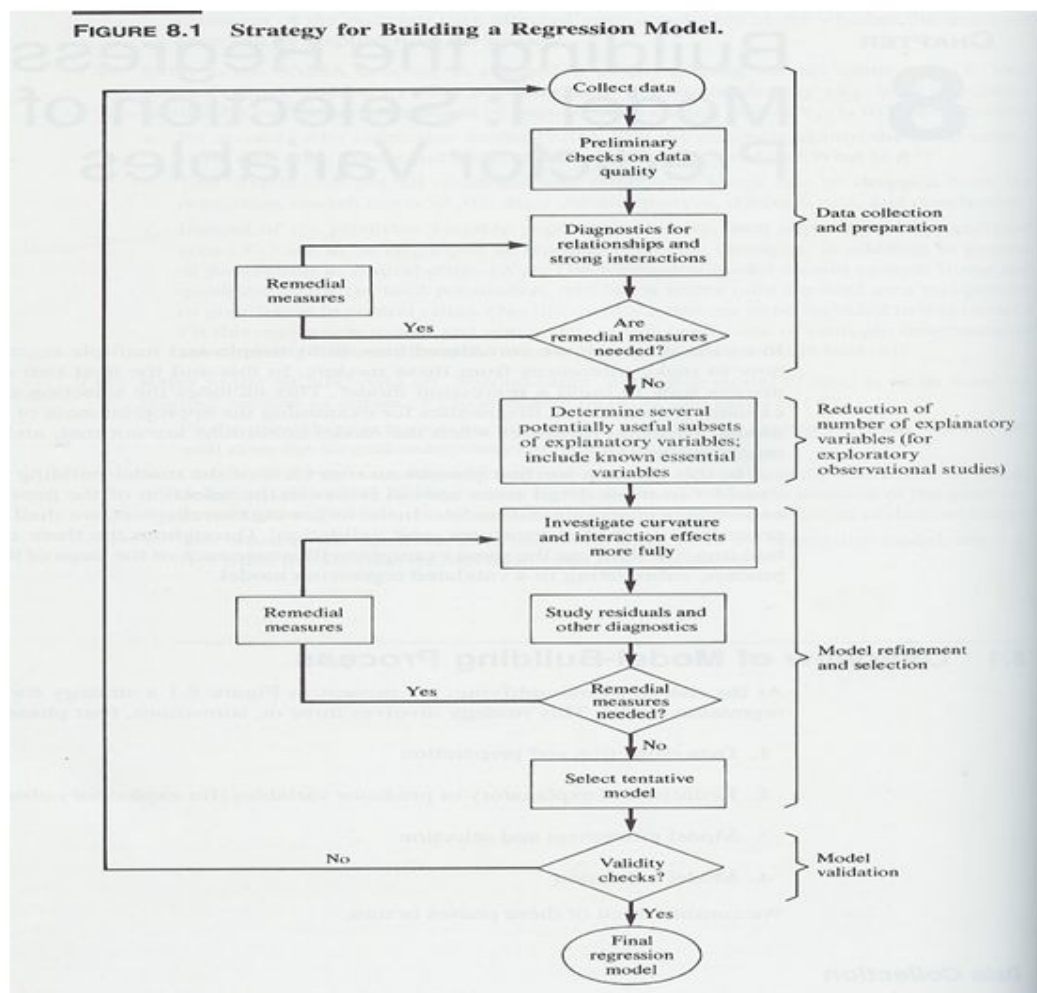
Pembentukan suatu model didasarkan pada tujuan dari si peneliti untuk apa model ini dibuat, apakah untuk keperluan prediksi ataukah untuk

menjelaskan hubungan di antara variabel-variabel yang diteliti. Oleh karena itu dalam proses pembuatannya dibutuhkan pengetahuan yang melatar belakangi studi tersebut baik secara teoritis maupun dari penelitian sebelumnya.

Strategi atau prosedur dalam pembentukan model regresi meliputi 4 tahap berikut ini:

1. Pengumpulan data dan persiapan (*Data collection and preparation*).
2. Pengurangan variabel bebas yang masuk ke dalam model
3. Penyempurnaan dan Seleksi model (*Model refinement and selection*)
4. Validasi model (*Model Validation*)

Secara ringkas tahapan strategi atau prosedur pembentukan model dapat ditampilkan dalam bentuk bagan berikut ini:



Sumber Gambar: *Applied Linear Statistical Models*, Kutner, Nachtsheim, Neter, Li.

Pengumpulan Data (*Data Collection*)

Persyaratan pengumpulan data untuk pembentukan model bervariasi tergantung dengan tipe studinya. Berikut ini 4 macam jenis studi/ penelitian:

1. *Controlled experiments*: Peneliti dapat mengendalikan level perlakuan (*treatment*) pada setiap unit percobaan.
2. *Controlled experiments with covariates*: dalam hal ini peneliti memasukkan variabel tidak terkontrol (*uncontrolled variables*) atau *covariates* ke dalam model.
3. *Confirmatory Observational Studies*: penelitian berdasarkan observasional bukan percobaan. Hipotesis penelitian disusun berdasarkan penelitian-penelitian sebelumnya.
4. *Explanatory Observational Studies*: biasanya dilakukan di bidang sosial, studi perilaku, ilmu-ilmu kesehatan, manajemen dll. Dalam penelitian ini si peneliti mencari variabel bebas yang mungkin berkaitan dengan variabel respon, karena tidak tersedianya informasi dari penelitian sebelumnya.

Persiapan Data (*Data Preparation*)

Jika proses pengumpulan data telah dilakukan, tahapan selanjutnya adalah persiapan data. Persiapan data di sini meliputi kegiatan *edit checks*, data error, deteksi pengamatan outlier atau pengamatan ekstrem.

Investigasi model awal (*Preliminary Model Investigation*)

Ketika data sudah siap untuk diolah dan sudah diperiksa dengan benar maka proses pembentukan model sudah dapat dilakukan.

1. Meneliti hubungan fungsional antara variabel bebas dengan variabel tidak bebasnya.
2. Mendeteksi apakah terdapat interaksi antar variabel bebasnya yang perlu dimasukkan ke dalam model.
3. Scatter plot dan residual plot sangat berguna untuk mengetahui kekuatan hubungan antara variabelnya.

4. Pengetahuan dari penelitian sebelumnya dan pendapat para ahli sangat penting dalam pembentukan model.

Pengurangan variabel bebas (*Reduction of Explanatory Variables*)

Hal yang paling sering dilakukan oleh peneliti adalah pertama-tama memasukkan semua variabel bebas ke dalam model. Kemudian membuang satu persatu variabel bebas yang tidak signifikan dalam model dengan melihat dari nilai statistik-t atau dari nilai p. value (signifikansinya).

Penyempurnaan dan seleksi Model (*Model refinement and selection*)

Pada tahap ini akan ada beberapa kandidat model yang bisa dipilih. Perlu diteliti secara detail tentang bentuk kurva responnya dan ada tidaknya pengaruh interaksi antar variabel bebasnya. Residual plot sangat berguna untuk memutuskan apakah model yang satu lebih bagus dari model lainnya. Cek diagnostik juga berguna dilakukan untuk mengidentifikasi pengamatan yang berpengaruh, atau untuk melihat adanya multikolinieritas atau tidak.

Jika kita menggunakan prosedur seleksi model yang otomatis dari software statistik, dan jika ditemukan satu model yang paling bagus, maka model lainnya juga masih bisa di eksplorasi lagi. Setelah mengecek semua asumsi dari analisis regresi terpenuhi atau tidak, maka kita dapat memutuskan model regresi final dan meneliti seberapa baik model yang kita pilih ini.

Criteria for Model Selection

Jika dari satu set model terdapat misalkan $p-1$ variabel bebas, maka alternatif model yang bisa dibuat sebanyak 2^{p-1} model. Misal, terdapat 3 variabel bebas dalam suatu model, maka akan terdapat $2^3 = 8$ kemungkinan model yang berbeda yang bisa dibentuk, yaitu :

1. Model tanpa variabel bebas : $y = b_0$
2. Model dengan satu variabel bebas

$$y = b_0 + b_1X_1$$

$$y = b_0 + b_2X_2$$

$$y = b_0 + b_3X_3$$

3. Model dengan dua variabel bebas:

$$y = b_0 + b_1X_1 + b_2X_2$$

$$y = b_0 + b_1X_1 + b_2X_3$$

$$y = b_0 + b_2X_2 + b_3X_3$$

4. Model dengan 3 variabel bebas: $y = b_0 + b_1X_1 + b_2X_2 + b_3X_3$

Tidak mungkin peneliti atau analis melakukan pengujian satu per satu untuk semua kemungkinan model yang ada, bayangkan jika terdapat 10 variabel bebas di dalam model, maka akan terdapat $2^{10} = 1024$ kemungkinan model regresi. Oleh karena itu kita menggunakan SPSS atau software statistik lainnya untuk melakukan proses seleksi ini.

Terdapat banyak ukuran yang bisa digunakan untuk mengukur model satu lebih bagus dari pada model lainnya dengan memperhatikan ukuran kriteria dibawah ini:

1. R_p^2 adalah nilai koefisien determinasi, model dikatakan bagus jika memiliki nilai R_p^2 yang tinggi, tetapi ada kecenderungan bahwa jika kita menambahkan variabel bebas ke dalam model, maka nilai koefisien determinasi juga naik.

2. R_{ap}^2 **adjusted** atau koefisien determinasi yang terkoreksi maksudnya ukuran ini memperhitungkan jumlah parameter dalam model, sehingga nilai koefisien determinasi terkoreksi oleh jumlah parameter yang ada di dalam model.

3. **C_p Mallow**, menggunakan kriteria ini artinya kita cari yang nilai C_p nya kecil dan mendekati p (banyaknya parameter dalam model).

4. **AIC_p (Akaike's Information Criterion)**, ini adalah kriteria seleksi model yang juga memberikan penalti pada model yang mempunyai banyak prediktor (variabel bebasnya). Kita pilih model yang mempunyai nilai AIC paling kecil.

Automatic Search Procedure for Model Selection

Sudah dijelaskan pada bagian sebelumnya bahwa jumlah kemungkinan model yang bisa dibuat dirumuskan dengan 2^{p-1} , sehingga semakin banyak jumlah variabel bebas dalam model semakin besar juga jumlah kemungkinan model yang dapat dibentuk. Untuk mempermudah kerja kita, maka *Automatic*

Search Computer sudah ada dalam paket program statistik. Sehingga kita bisa langsung menggunakannya. Ada beberapa prosedur yang bisa dipakai, diantaranya:

1. “Best” Subsets Algorithms
2. Stepwise Regression Methods
3. Forward Stepwise Regression
4. Forward Selection
5. Backward Elimination

Validasi Model (*Model Validation*)

Validitas model adalah bentuk kestabilan dan kewajaran koefisien regresi. Dapat diterima secara nalar dan model regresi dapat dipakai untuk membuat kesimpulan secara umum. Validasi adalah bagian yang diperlukan dalam proses pembuatan model. Ini adalah tahap akhir dari *model building process*. Ada 3 cara dalam validasi model:

1. Pengumpulan data baru
2. Perbandingan hasil antara hasil awal dengan hasil simulasi.
3. Gunakan data sampel lainnya untuk mengecek model dan kemampuan prediktifnya.

Contoh Soal:

Bagian/unit operasi sebuah rumah sakit tertarik untuk memprediksi “survival” dari pasien yang telah melakukan operasi liver. Dipilih sampel acak sebanyak 54 pasien. Setiap pasien disimpan data rekam medisnya untuk dievaluasi. Berikut catatan data rekam medis pasien yang berhasil dikumpulkan.

Y = Survival time

X_1 = blood clotting score (skore penggumpalan darah)

X_2 = Prognostic index

X_3 = Enzyme function test score

X_4 = Liver function test score

X_5 = Age (years)

X_6 = variabel indikator jenis kelamin (0=laki-laki; 1 = perempuan)

X_7 dan X_8 = variabel indikator riwayat konsumsi alkohol

Konsumsi Alkohol	X ₇	X ₈
Tidak mengonsumsi	0	0
Sedang	1	0
Berat	0	1

Data selengkapnya dapat diunduh di:

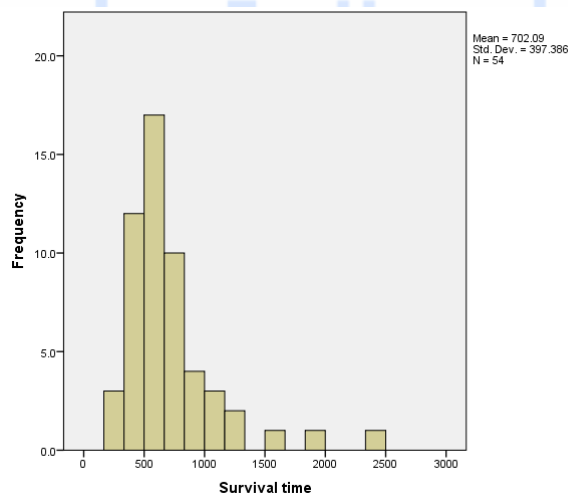
<http://www.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerData/Chapter%20%209%20Data%20Sets/CH09TA01.txt>

Berdasarkan data tersebut: Untuk latihan awal dibatasi sampai 4 variabel pertama saja, yaitu variabel X1, X2, X3 dan X4.

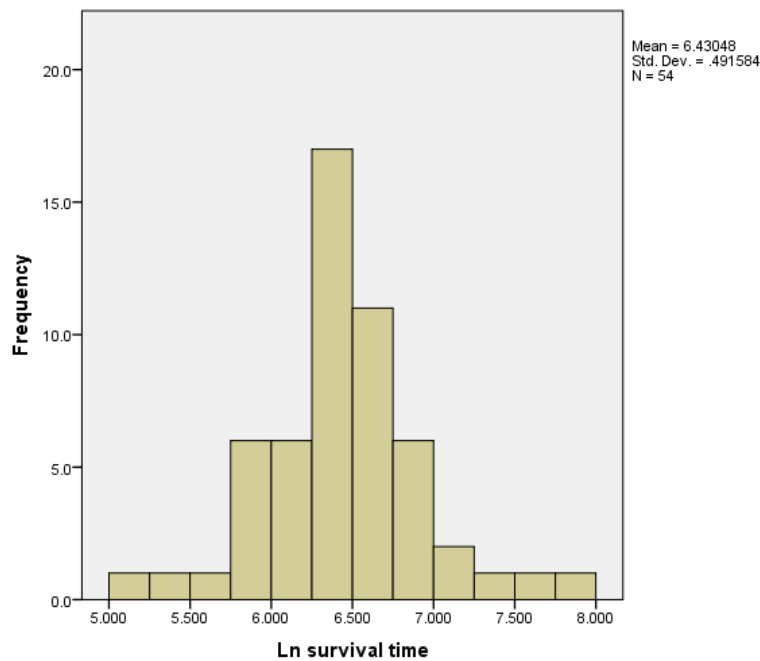
- Lakukan explorasi data, lihat sebaran variabel responnya dan buat matriks korelasi antar variabel!
- Berdasarkan data rekam medis pasien, maka buatlah model terbaik untuk memprediksi time survival pasien di rumah sakit tersebut.

Jawab:

- Explorasi data penting dilakukan untuk mengetahui gambaran data yang akan kita analisis. Salah satu asumsi analisis regresi adalah Y menyebar mengikuti distribusi normal.

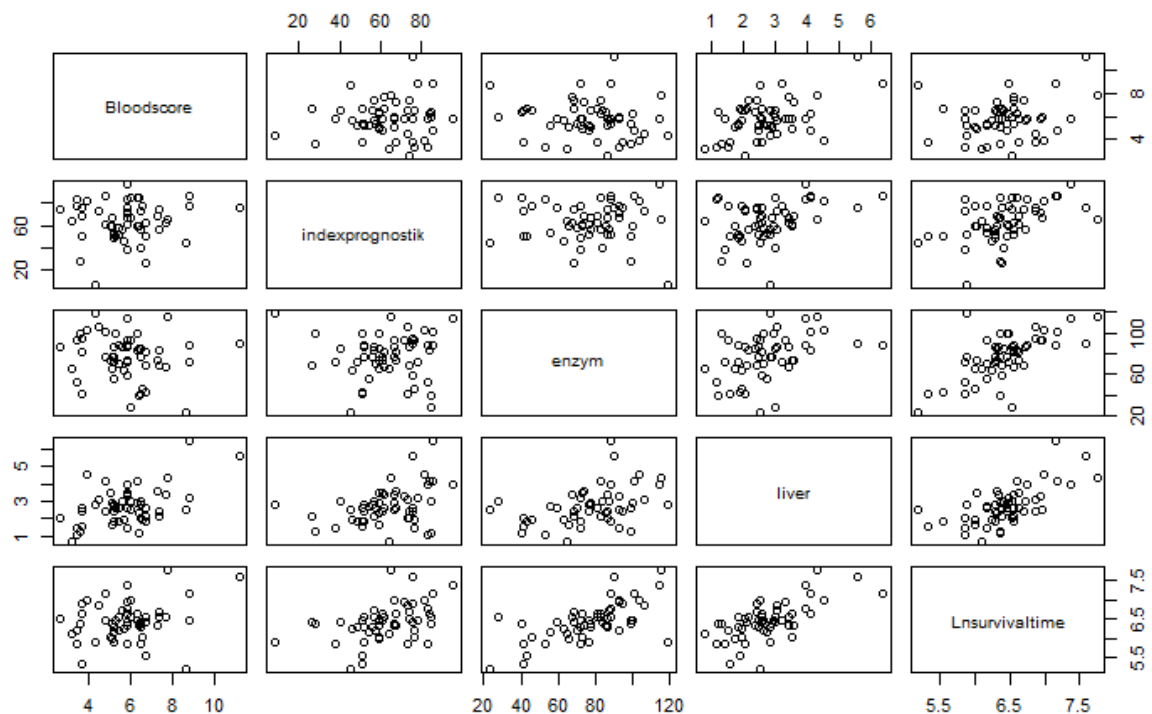


Salah satu cara mengetahui menyebar normal atau tidak dengan membuat histogram untuk variabel tidak bebasnya, terlihat bahwa bentuk sebarannya cenderung miring (tidak normal), sehingga perlu dilakukan transformasi kepada variabel responnya. Transformasi yang sering dipakai adalah $y' = \ln Y$.



Setelah dilakukan transformasi terlihat indikasi normal. Sehingga untuk analisis berikutnya variabel tidak bebas (Y) diganti menjadi $Y' = \ln Y$.

Matriks korelasi sangat berguna untuk mengetahui pola hubungan antar variabelnya.



b. Model terbaik dalam kasus ini coba didekati dengan menggunakan SPSS dan memilih metode stepwise, diperoleh hasil sebagai berikut:

Dari hasil output SPSS diperoleh bahwa model terbaik yang bisa digunakan adalah model dengan melibatkan variabel bebas X_1 (Blood score), X_2 (prognostic index) dan X_3 (enzymes score test), karena memiliki nilai $R_p^2 = 0.757$ dan $R_{ap}^2 = 0.743$ paling besar.

Regression

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	enzymes score test		Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	index prognostik		Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
3	Blood score		Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: Ln survival time

Model Summary^d

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.654 ^a	.428	.417	.375489
2	.814 ^b	.663	.650	.290790
3	.870 ^c	.757	.743	.249341

a. Predictors: (Constant), enzymes score test

b. Predictors: (Constant), enzymes score test, index prognostik

c. Predictors: (Constant), enzymes score test, index prognostik, Blood score

d. Dependent Variable: Ln survival time

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5.476	1	5.476	38.840	.000 ^b
	Residual	7.332	52	.141		
	Total	12.808	53			
2	Regression	8.495	2	4.248	50.233	.000 ^c
	Residual	4.312	51	.085		
	Total	12.808	53			
3	Regression	9.699	3	3.233	52.003	.000 ^d
	Residual	3.109	50	.062		
	Total	12.808	53			

a. Dependent Variable: Ln survival time

b. Predictors: (Constant), enzymes score test

c. Predictors: (Constant), enzymes score test, index prognostik

d. Predictors: (Constant), enzymes score test, index prognostik, Blood score

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5.264	.194		27.138	.000
	enzymes score test	.015	.002	.654	6.232	.000
2	(Constant)	4.351	.214		20.296	.000
	enzymes score test	.015	.002	.665	8.186	.000
	index prognostik	.014	.002	.486	5.975	.000
3	(Constant)	3.766	.227		16.609	.000
	enzymes score test	.016	.002	.711	10.094	.000
	index prognostik	.013	.002	.459	6.556	.000
	Blood score	.095	.022	.311	4.401	.000

a. Dependent Variable: Ln survival time

Excluded Variables^a

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	Blood score	.352 ^b	3.699	.001	.460	.978
	index prognostik	.486 ^b	5.975	.000	.642	.999
	liver score test	.456 ^b	4.679	.000	.548	.827
2	Blood score	.311 ^c	4.401	.000	.528	.970
	liver score test	.283 ^c	3.108	.003	.402	.683
3	liver score test	.070 ^d	.625	.535	.089	.391

a. Dependent Variable: Ln survival time

b. Predictors in the Model: (Constant), enzymes score test

c. Predictors in the Model: (Constant), enzymes score test, index prognostik

d. Predictors in the Model: (Constant), enzymes score test, index prognostik, Blood score

Residuals Statistics^a

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	5.47423	7.47597	6.43048	.427789	54
Residual	-.469920	.592526	.000000	.242181	54
Std. Predicted Value	-2.235	2.444	.000	1.000	54
Std. Residual	-1.885	2.376	.000	.971	54

a. Dependent Variable: Ln survival time



Materi 5

Diagnostik Model Regresi

Pendahuluan

Setelah melakukan model building, langkah selanjutnya adalah mendiagnosa model yang didapat apakah sudah layak atau bagus tidak untuk digunakan untuk prediksi maupun pemodelan. Diagnosa yang dimaksud di sini adalah mahasiswa mampu mendiagnosa pengamatan yang mempengaruhi model regresi. Selain itu mahasiswa mampu menguraikan tentang pengamatan ekstrem (outlier) dan pengamatan yang berpengaruh (*influential cases*) terhadap model regresi.

Diagnostik Model Regresi

Diagnosa yang dilakukan pada model regresi yang sudah kita dapatkan adalah sebagai berikut:

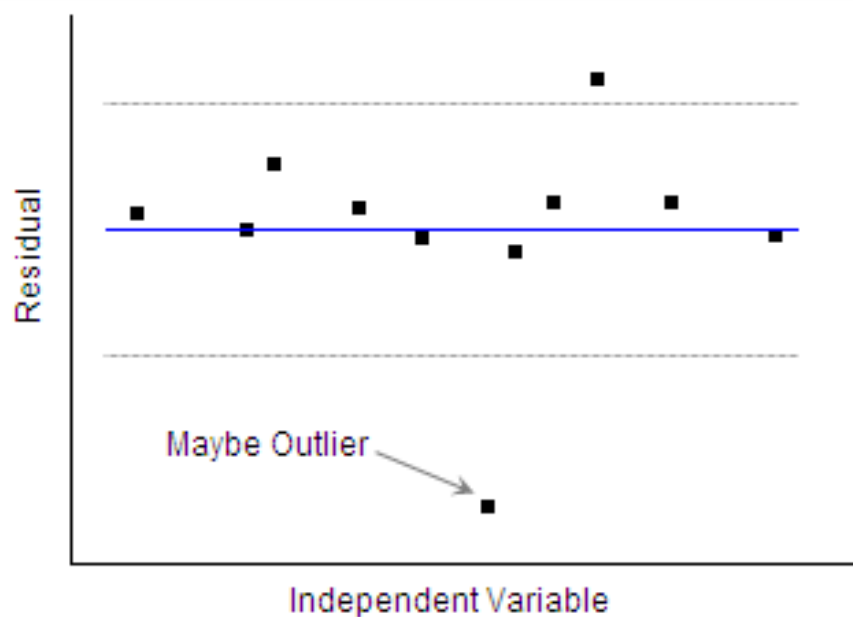
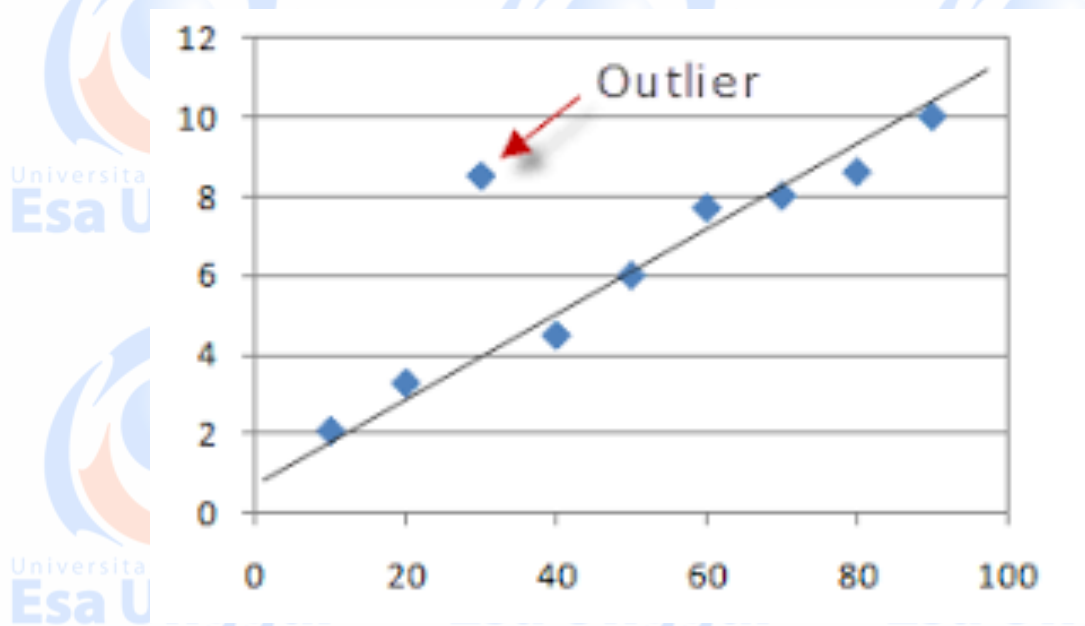
1. Pengamatan tidak biasa dan pengamatan yang berpengaruh (*Unusual and influential data*):

- **Outlier:** adalah suatu observasi yang menyimpang dari sekumpulan data yang lain atau pengamatan yang jauh dari pusat data. Nilai Y (variabel tidak bebas) menjadi sangat tidak biasa ketika diberikan nilai X- tertentu. Outlier memberikan indikasi adanya keanehan sampel atau bisa juga indikasi dari kesalahan entry data.
- **Leverage** adalah pengamatan dengan nilai ekstrim pada variabel bebas (prediktor). Leverage mengukur seberapa jauh nilai variabel Y menyimpang dari nilai rata-ratanya. Leverage ini bisa mempengaruhi estimasi koefisien regresinya.
- **Influence point:** adalah perpaduan antara outlier dan leverage, artinya jika kita menghilangkan pengamatan ini maka estimasi koefisien regresi akan sangat berubah.

2. Heterokedasitas (varians tidak konstan)

3. Multikolinieritas (terdapat hubungan antar variabel bebas atau *non independence of x-variables*).

4. Autokorelasi.



Keberadaan data outlier/ pencilan akan mengganggu dalam proses analisis data dan harus dihindari, karena dapat menyebabkan error dari model semakin besar, varians pada data juga menjadi besar dan estimasi interval dari koefisien regresi juga semakin lebar.

Berikut ini ukuran untuk mendeteksi outlier

1. Leverage: menampilkan nilai leverage (pengaruh) terpusat.
 2. DFFITS atau Standardized DFFIT: menampilkan nilai perubahan yang diprediksi jika case tertentu dikeluarkan, yang sudah distandarkan.
 3. Cook's Distance: menampilkan nilai jarak Cook
 4. DFBETA(s) : Menampilkan nilai perubahan koefisien regresi hasil perubahan yang disebabkan oleh pengeluaran case tertentu.
- Digunakan untuk mendeteksi outlier pada variabel bebas.

Ketentuan dalam pengambilan keputusan:

1. Leverage $> (2p-1)/n$, artinya terdapat outlier/ pencilan
 2. DFFITS $> 2 \cdot \sqrt{p/n}$, artinya terdapat outlier/ pencilan
 3. Cook's Distance $> F(0.5; p, n-p)$, artinya terdapat outlier/ pencilan
 4. DFBETA(s) $> 2/\sqrt{n}$, artinya terdapat outlier/ pencilan
- Dimana n = jumlah observasi (sampel) dan p = jumlah parameter dalam model.

Materi 6

Uji Asumsi Analisis Regresi Linier Berganda

Pendahuluan

Pada perkuliahan sebelumnya sudah dilakukan proses model seleksi dan diagnosa model. Maka langkah selanjutnya adalah menguji apakah asumsi yang diperlukan dalam analisis regresi sudah dipenuhi atau tidak. Dalam bagian ini akan diberikan penjelasan mengenai apa saja yang menjadi landasan atau asumsi dari analisis regresi, kemudian jika ada salah satu asumsi yang tidak terpenuhi bagaimana cara mengatasinya. Uji asumsi analisis regresi berganda yang akan dipelajari adalah uji normalitas, uji heteroskedastisitas, uji multikolinieritas, dan uji autokorelasi.

Kompetensi Dasar

Mahasiswa mampu menguraikan apa yang menjadi asumsi analisis regresi linier berganda dan cara mengatasinya jika asumsi tersebut tidak terpenuhi.

Kemampuan Akhir yang Diharapkan

Mahasiswa dapat menganalisis ketepatan penggunaan analisis regresi dan korelasi berganda berdasarkan hasil dari uji asumsi dan standarisasi.

Kegiatan Belajar

Uji Asumsi Analisis Regresi Linier Berganda

Uraian dan contoh

Pendahuluan

Uji asumsi analisis regresi berganda perlu dilakukan sebagai salah satu syarat agar model regresi yang kita peroleh dapat dikatakan baik. Terdapat empat uji asumsi analisis regresi berganda yang harus dipenuhi, yaitu:

1. Uji Normalitas

Asumsi normalitas pada analisis regresi berganda tidak pada variabel, melainkan pada residualnya. Apabila tidak normal, maka dilakukan transformasi untuk membuat jadi normal. Jika gagal dapat diteliti apakah terdapat outlier atau menambah sampel.

Langkah-langkah Uji Normalitas:

Hipotesis:

H_0 : Residual menyebar normal

H_1 : Residual tidak menyebar normal

Untuk mengetahui apakah distribusi data pada setiap variabel terdistribusi secara normal atau tidak caranya dengan menggunakan Normal P-P plot, jika sebaran data berada di sekitar garis diagonal dan mengikuti arah diagonal, maka model regresi dikatakan memenuhi asumsi normalitas. Jika sebaran data tidak di sekitar garis diagonal atau menyebar berjauhan atau acak, maka data tidak menyebar normal atau model regresi tidak memenuhi asumsi normalitas.

Diagram Normal P-P plot dikatakan memenuhi asumsi normalitas jika diagram menunjukkan plot-plot mengikuti alur garis lurus.

Uji Normalitas dengan SPSS:

- Dari menu utama SPSS, pilih *analyze*
- Pilih sub menu *regression*
- Pilih *Linear*
- Isi *dependet dan independent variable* sesuai dengan kasus yang dihadapi
- Pilih kolom *plot*, klik *normal probability plot*
- Klik *continue*
- Klik *OK*

4. Uji Heteroskedastisitas:

Untuk melihat apakah terdapat ragam atau varians yang sama dari nilai residual satu pengamatan ke pengamatan lainnya. Agar analisis regresi berganda bisa dilakukan syaratnya harus memiliki ragam yang sama (homoskedastisitas) atau varians yang sama. Artinya setiap Y yang

berhubungan dengan berbagai nilai X mempunyai varians yang sama. Jika varians tidak sama maka disebut dengan istilah heteroskedastisitas

Hipotesis:

H_0 : Tidak terjadi heteroskedastisitas

H_1 : Terjadi heteroskedastisitas

Cara mendeteksinya dengan menggunakan scatter plot, jika membentuk suatu pola aka terjadi heteroskedastisitas, apabila tidak membentuk suatu pola tertentu maka tidak terjadi heteroskedastisitas.

Model regresi yang memenuhi syarat adalah dimana terdapat kesamaan varians dari residual satu pengamatan ke pengamatan yang lain tetap atau disebut homoskedastisitas.

Deteksi heteroskedastisitas dapat dilakukan dengan metode scatter plot dengan memplotkan nilai ZPRED (nilai Prediksi) dengan SRESID (nilai residualnya). Tidak terjadi heteroskedastisitas jika tidak terdapat pola tertentu pada grafik (tidak ada pengumpulan data di tengah menyerupai bulatan, menyempit atau melebar menyerupai corong). Solusi jika model tidak memenuhi asumsi adalah dengan mentransformasi ke dalam bentuk logaritma. Kecenderungan semakin membesar varian akan mengakibatkan uji hipotesis yang dilakukan tidak akan memberikan hasil yang baik (tidak valid).

Uji Heteroskedastisitas dengan SPSS:

- Dari menu utama SPSS, pilih *analyze*
- Pilih sub menu *regression*
- Pilih *Linear*
- Isi *dependet dan independent variable* sesuai dengan kasus yang dihadapi
- Pilih kolom *plot*, pada bagian Y klik ZRESID, pada bagian X klik ZPRED
- Klik *continue*
- Klik OK

5. Uji Multikolinearitas

Multikolinearitas adalah suatu kondisi dimana terdapat hubungan yang kuat antar variabel bebasnya. Kondisi yang harus dipenuhi dalam analisis regresi berganda adalah tidak terjadi multikolinearitas. Suatu data terbebas dari masalah multikolinearitas jika:

- VIF (Variance Inflation Factor) < 10
- Tolerance > 0.1

Uji Multikolinearitas dengan SPSS:

- Dari menu utama SPSS, pilih *analyze*
- Pilih sub menu *regression*
- Pilih *Linear*
- Isi *dependet dan independent variable* sesuai dengan kasus yang dihadapi
- Pilih kolom *statistic, pilih estimates, covariance matrix, collinearity diagnostics dan model fit*
- Klik *continue*
- Klik *OK*

6. Uji Autokorelasi

Autokorelasi, artinya adanya korelasi berurutan, dimana jika terdapat dua pengamatan yang berbeda maka saling berkorelasi. Model regresi yang baik mensyaratkan tidak adanya masalah autokorelasi.

Hipotesis:

H_0 : Tidak terjadi autokorelasi

H_1 : terjadi autokorelasi

Dengan cara membandingkan nilai statistik uji durbin watson (d) dengan tabel durbin watson (nilai dL dan nilai dU).

Keputusan:

$dU < d < 4 - dU$ maka H_0 diterima (tidak terjadi autokorelasi)

$d < dL$ atau $d > 4 - dL$ maka H_0 ditolak (terjadi autokorelasi)

$dL < d < dU$ atau $4 - dU < d < 4 - dL$, maka tidak ada kesimpulan.

Uji Autokorelasi dengan SPSS:

- Dari menu utama SPSS, pilih *analyze*
- Pilih sub menu *regression*
- Pilih *Linear*
- Isi *dependet dan independent variable* sesuai dengan kasus yang dihadapi
- Pilih kolom *statistic*, pilih *D-W*
- Klik *continue*
- Klik *OK*

Menurut teorema Gauss-Markov, setiap estimator OLS harus memenuhi kriteria BLUE yaitu: Best, Linear, Unbiased (Gujarati, 1995: 72-72).

Best = yang terbaik, hasil regresi dikatakan “best” apabila garis regresi yang dihasilkan untuk melakukan estimasi atau peramalan dari sebaran data mempunyai error yang kecil.

Linear = merupakan kombinasi dari data sampel

Linear dalam model artinya model yang digunakan dalam analisis regresi telah sesuai dengan kaidah OLS (*Ordinary Least Square*) dimana variabel-variabel penduganya hanya berpangkat satu. Sedangkan linear dalam parameter menjelaskan bahwa parameter yang dihasilkan merupakan fungsi linear dari sampel.

Unbiased = rata-rata nilai harapan harus sama dengan nilai sebenarnya atau dengan kata lain tidak bias. Estimator dikatakan tidak bias jika nilai harapan estimator b sama dengan nilai yang benar dari b (nilai rata-rata $b = b$). Jika nilai rata-rata b tidak sama dengan nilai b maka selisishnya itu disebut bias.

Contoh Soal dan Penyelesaian

Seorang manajer sebuah rumah sakit ingin mengetahui profit rumah sakit yang dipimpinnya, apakah total profit yang diperoleh dipengaruhi oleh besarnya biaya peralatan dan biaya pegawai rumah sakit. Berikut data yang dikumpulkan selama 10 tahun terakhir.

Biaya Perlengkapan			
Tahun	Rumah Sakit	Biaya Pegawai	Profit
1	6350	1250	10150
2	6540	1200	12540
3	6798	1350	13850
4	7230	1350	14650
5	7625	1425	16400
6	7750	1360	18700
7	6780	1200	19500
8	7270	1450	19675
9	7450	1750	21450
10	8400	1795	22500

Analisa dengan menggunakan tingkat keyakinan 95%, apakah model regresi tersebut layak digunakan? Ujilah juga asumsi analisis regresi berganda!

Jawab:

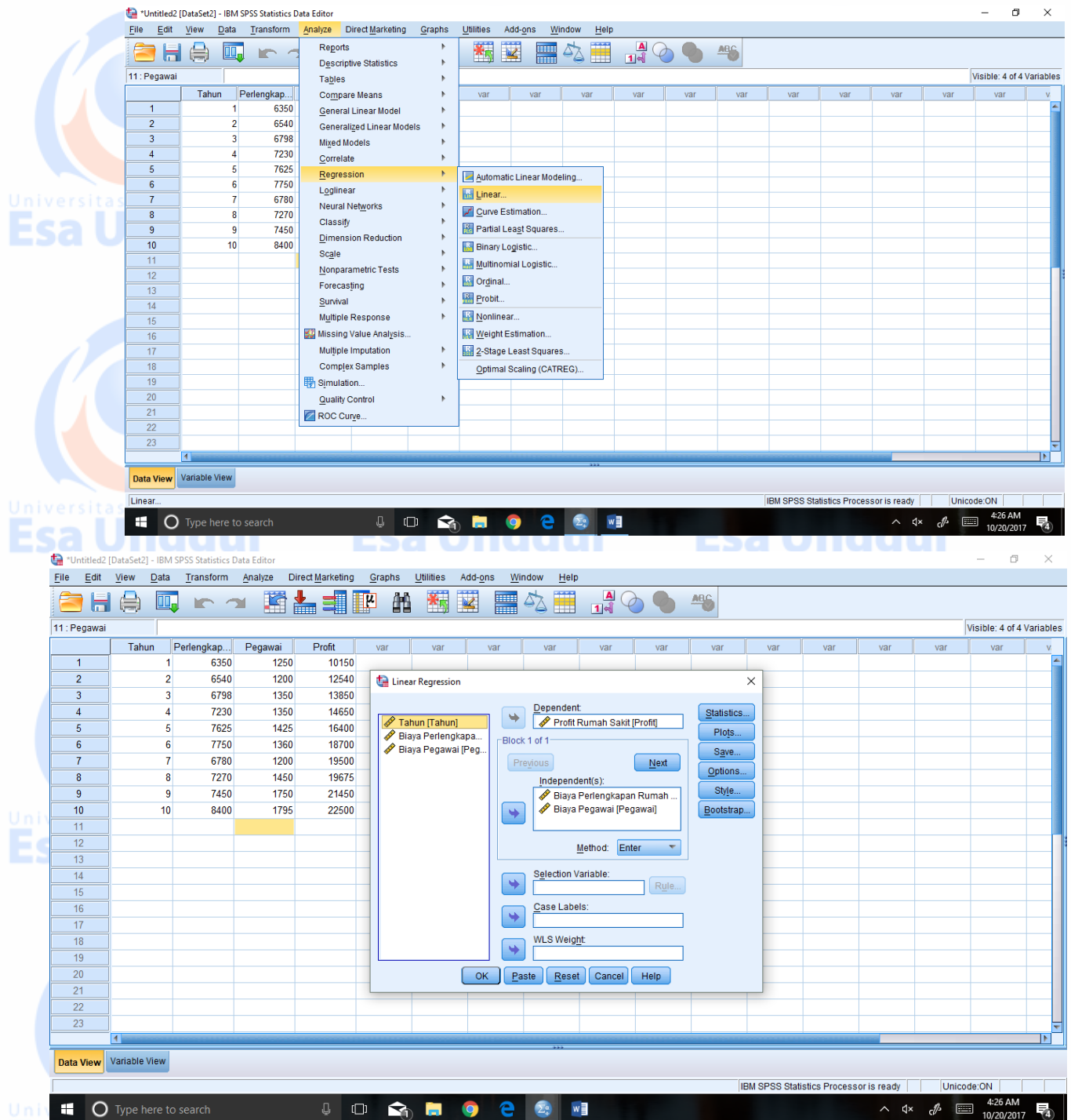
Berikut ini adalah output SPSS dari hasil analisis regresi, dimana:

Y = Profit rumah sakit

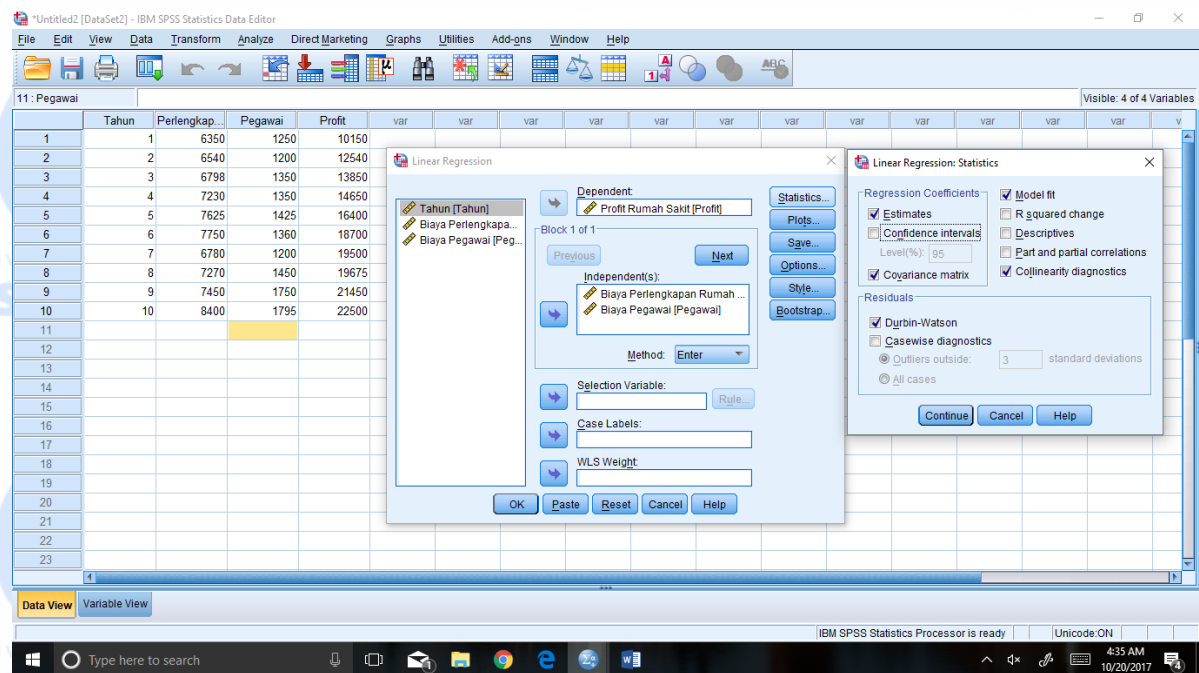
X_1 = Biaya perlengkapan rumah sakit

X_2 = Biaya pegawai

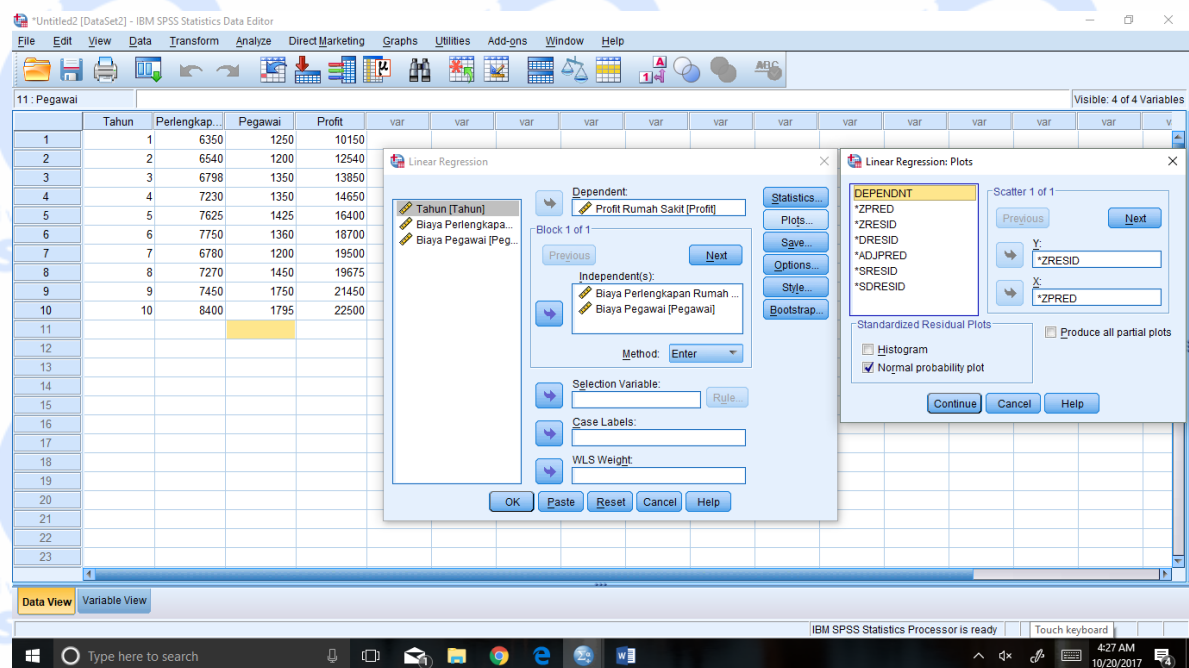
Berikut ini cara melakukan analisis regresi berganda dan uji asumsinya melalui SPSS dengan data di atas:



Untuk melakukan estimasi parameter dan fit model regresi berganda maka pada kolom *statistics*, pilih menu *estimates* dan *model fit*, sedangkan untuk mengetahui terjadi atau tidaknya autokorelasi klik *Durbin-Watson* serta klik *covariance matrix* untuk melihat hubungan antar variabelnya.



Untuk uji normalitas dan menampilkan normal probability plot lakukan langkah berikut:



Dari langkah-langkah di atas, maka akan diperoleh output SPSS sebagai berikut:

Regression

Notes

Output Created	20-OCT-2017 04:36:50	
Comments		
Input	Active Dataset	DataSet2
	Filter	<none>
	Weight	<none>
	Split File	<none>
	N of Rows in Working Data File	10
Missing Value Handling	Definition of Missing	User-defined missing values are treated as missing.
	Cases Used	Statistics are based on cases with no missing values for any variable used.
Syntax	REGRESSION /MISSING LISTWISE /STATISTICS COEFF OUTS BCOV R ANOVA COLLIN TOL /CRITERIA=PIN(.05) POUT(.10) /NOORIGIN /DEPENDENT Profit /METHOD=ENTER Perlengkapan Pegawai /SCATTERPLOT=(*ZRESID ,*ZPRED) /RESIDUALS DURBIN NORMPROB(ZRESID).	
Resources	Processor Time	00:00:00.22
	Elapsed Time	00:00:00.23
	Memory Required	1660 bytes
	Additional Memory Required for Residual Plots	560 bytes

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	Biaya Pegawai, Biaya Perlengkapan Rumah Sakit ^b		Enter

a. Dependent Variable: Profit Rumah Sakit

b. All requested variables entered.

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.791 ^a	.626	.520	2819.515	.931

a. Predictors: (Constant), Biaya Pegawai, Biaya Perlengkapan Rumah Sakit

b. Dependent Variable: Profit Rumah Sakit

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	93285346.717	2	46642673.359	5.867	.032 ^b
	Residual	55647655.783	7	7949665.112		
	Total	148933002.500	9			

a. Dependent Variable: Profit Rumah Sakit

b. Predictors: (Constant), Biaya Pegawai, Biaya Perlengkapan Rumah Sakit

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-16462.974	11486.086		-1.433	.195		
	Biaya Perlengkapan Rumah Sakit	3.457	2.435	.529	1.419	.199	.385	2.599
	Biaya Pegawai	5.978	7.288	.306	.820	.439	.385	2.599

a. Dependent Variable: Profit Rumah Sakit

Coefficient Correlations^a

Model			Biaya Pegawai	Biaya Perlengkapan Rumah Sakit
1	Correlations	Biaya Pegawai	1.000	-.784
		Biaya Perlengkapan Rumah Sakit	-.784	1.000
	Covariances	Biaya Pegawai	53.112	-13.922
		Biaya Perlengkapan Rumah Sakit	-13.922	5.931

a. Dependent Variable: Profit Rumah Sakit

Collinearity Diagnostics^a

Model		Dimension	Eigenvalue	Condition Index	Variance Proportions		
					(Constant)	Biaya Perlengkapan Rumah Sakit	Biaya Pegawai
1	1		2.989	1.000	.00	.00	.00
	2		.010	17.621	.29	.00	.40
	3		.002	41.963	.71	1.00	.59

a. Dependent Variable: Profit Rumah Sakit

Residuals Statistics^a

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	12961.86	23306.86	16941.50	3219.478	10
Residual	-2811.863	5350.507	.000	2486.579	10
Std. Predicted Value	-1.236	1.977	.000	1.000	10
Std. Residual	-.997	1.898	.000	.882	10

a. Dependent Variable: Profit Rumah Sakit



Universitas
Esa Unggul



Universitas
Esa Unggul

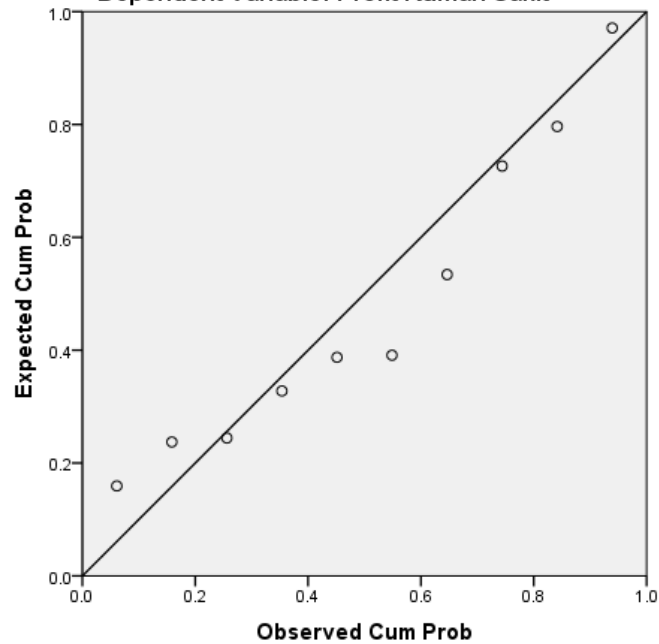


Universitas
Esa Unggul

Charts

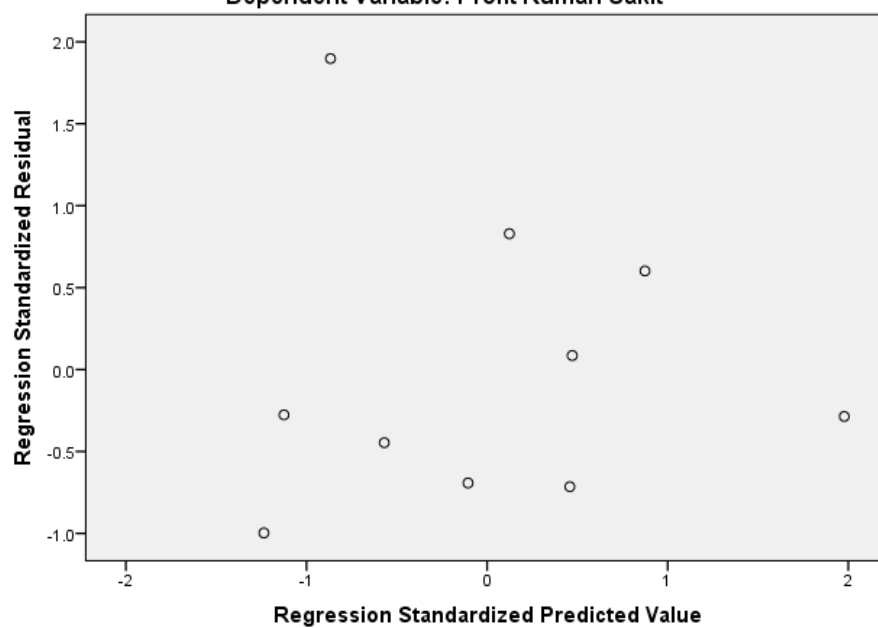
Normal P-P Plot of Regression Standardized Residual

Dependent Variable: Profit Rumah Sakit



Scatterplot

Dependent Variable: Profit Rumah Sakit



JAWABAN UJI ASUMSI ANALISIS REGRESI BERGANDA:

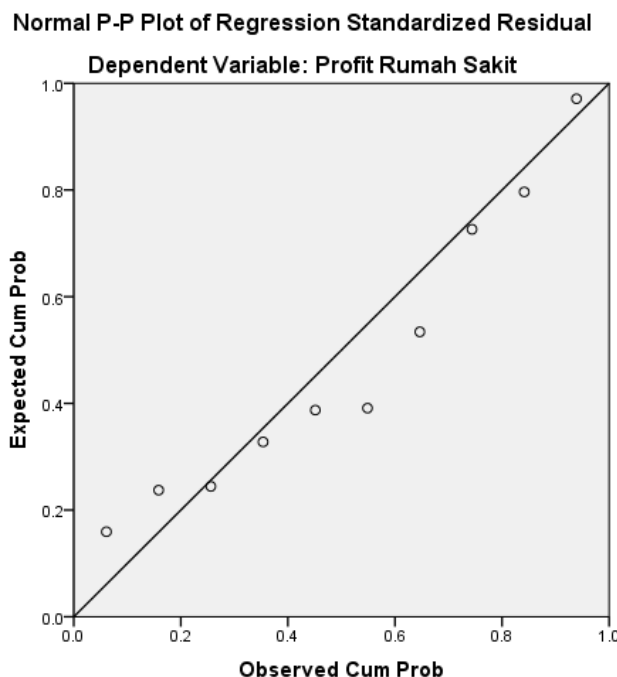
1. Uji Normalitas

H_0 : Residual menyebar normal

H_1 : Residual tidak menyebar normal

Lihat hasil dari grafil Normal P-P Plot di bawah ini, terlihat residual menyebar di sekitar garis diagonal, sehingga kita menerima H_0 yang artinya data terdistribusi secara normal dengan tingkat keyakinan 95%.

Sebenarnya dengan hanya melihat grafik Normal P-P Plot masih bersifat agak subyektif, maka diperlukan suatu ukuran yang akurat yang menampilkan nilai signifikansi sehingga kita dapat menyimpulkan normal atau tidak dengan uji *Kolmogorov-Smirnov*.



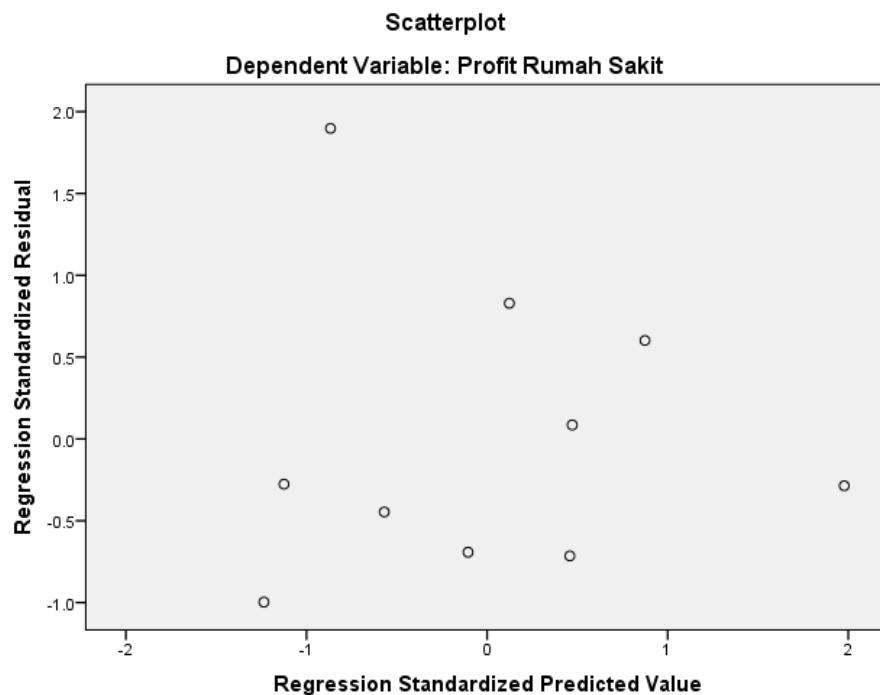
2. Uji Heteroskedastisitas

H_0 : Varians sama (Homoskedastisitas)

H_1 : Varians berbeda (Heteroskedastisitas)

Perhatikan output SPSS pada *scatterplot* antara ZPREDICT dengan ZRESID di bawah ini. Kalau kita perhatikan tidak membentuk pola tertentu sehingga

kita dapat menerima H_0 yang artinya tidak terjadi heteroskedastisitas, sehingga model regresi yang baik dapat terpenuhi.



3. Uji Multikolinearitas

Dari output SPSS lihat tabel *coefficient* yang mengandung nilai VIF

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1 (Constant)	-16462.974	11486.086		-1.433	.195		
Biaya Perlengkapan Rumah Sakit	3.457	2.435	.529	1.419	.199	.385	2.599
Biaya Pegawai	5.978	7.288	.306	.820	.439	.385	2.599

a. Dependent Variable: Profit Rumah Sakit

VIF 2,599 < 10 dan Tolerance 0,385 > 0,1 maka terima H_0 artinya tidak terdapat multikolinearitas dengan tingkat keyakinan 95%.

4. Uji Autokorelasi

H_0 : Tidak terjadi autokorelasi

H_1 : terjadi autokorelasi

Dari output SPSS lihat tabel *model summary* yang mengandung nilai *Durbin Watson*.

Model Summary^a

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.791 ^a	.626	.520	2819.515	.931

a. Predictors: (Constant), Biaya Pegawai, Biaya Perlengkapan Rumah Sakit

b. Dependent Variable: Profit Rumah Sakit

Diketahui:

N = 10 (jumlah data)

k = 3 (jumlah parameter dalam model regresi yaitu b_0 , b_1 , dan b_2)

Durbin-Watson = $d = 0,931$

Dibandingkan dengan nilai di tabel Durbin-Watson dengan $\alpha = 5\%$ dengan $k=3$ dan $n=10$, diperoleh:

$dL = 0,5948$ dan $dU = 1,928$

Nilai Durbin-Watson (d) berada diantara nilai dL dan dU ($dL < d < dU$) = ($0,5948 < d < 1,928$) maka tidak ada kesimpulan.

Materi 7

Student's Discussion Project

Pendahuluan

Pada pertemuan ke 7 ini, mahasiswa akan dilatih untuk memecahkan suatu kasus yang diberikan secara team. Akan diajarkan bagaimana mereka menaungkan hasil analisisnya dalam bentuk report atau laporan yang sistematis.

Kemampuan Akhir Mahasiswa yang diharapkan:

- Mahasiswa mampu menguasai semua materi yang sudah diajarkan dari pertemuan-1 sampai dengan pertemuan 6.
- Mengaplikasikan materi yang sudah diterangkan dalam sebuah project.
- Mahasiswa berlatih untuk mengeksplorasi, menganalisa serta membuat report atau laporan tertulis dan mempresentasikannya di depan kelas.

KASUS

- Data yang digunakan Patient Satisfaction
- Kepala bagian administrasi sebuah rumah sakit ingin mempelajari hubungan antara tingkat kepuasan pasien dengan umur pasien (tahun), tingkat keparahan penyakit (severity indeks) dan tingkat kecemasan pasien (anxiety index). Secara random dipilih 46 pasien untuk diteliti dan diperoleh data kepuasan pasien (patients satisfaction data).
- Data selengkapnya dapat diunduh di:
<http://www.stat.ufl.edu/~randles/sta4210/Rclassnotes/data/textdataset/s/KutnerData/Chapter%20%209%20Data%20Sets/CH09TA01.txt>

Lakukan langkah-langkah berikut ini:

1. Lakukan explorasi data untuk mendapatkan gambaran tentang perilaku data, dengan membuat deskriptif data dan matriks korelasi antar variabel.
2. Lakukan analisis regresi berganda dengan memasukkan semua variabel bebasnya
3. Tulislah persamaan regresi yang didapat pada langkah-2
4. Jelaskan arti dari koefisien regresi yang didapat
5. Berapakah koefisien korelasi dan koefisien determinasi bergandanya?
6. Apakah pengaruh dari masing-masing variabel bebasnya signifikan?
7. Apakah model regresi yang diperoleh layak digunakan untuk prediksi?
8. Lakukan model selection untuk memilih alternatif model lain yang lebih bagus.

Buatlah satu laporan atau report dengan format sebagai berikut:

1. Pendahuluan
2. Tinjauan Pustaka
3. Data dan Metode
4. Hasil dan Pembahasan
5. Kesimpulan dan Saran

Materi 8

Pengenalan Analisa Data Kategorik

Pada materi ini diharapkan mahasiswa mampu memahami konsep data kategorik, Odds Ratio (OR) dan Relative Risk (RR), dimana semua ukuran ini merupakan faktor resiko.

Faktor Resiko

Beberapa ukuran yang dapat digunakan untuk melihat faktor resiko diantaranya:

1. Relative Risk (RR)
2. Odds Ratio (OR)

Relative Risk (RR) dan Odds Ratio (OR) dipakai dalam studi epidemiologi untuk menjelaskan apakah ada hubungan antara variabel independen dengan variabel dependen atau ratio antara dua proporsi.

- RR biasanya dipakai untuk penelitian prospektif/ Kohort
- OR biasanya dipakai untuk penelitian retrospektif/ studi kasus kontrol

Paparan	Penyakit (+)	Penyakit (-)	Total
Positif	a	b	m1
Negatif	c	d	m2
Total	n1	n2	t

RR = rasio resiko = rasio relatif

RR = rasio p_1/p_2

p_1 (proporsi faktor resiko positif)

p_2 (proporsi faktor resiko negatif)

Dimana:

$p_1 = a/m_1$ (proporsi faktor resiko positif)

$p_2 = c/m_2$ (proporsi faktor resiko negatif)

- Untuk studi epidemiologi yang bertujuan untuk membandingkan kelompok terpapar dengan kelompok tidak terpapar maka RR merupakan indikator terbaik untuk menunjukkan hubungan antar keduanya.

- Jika untuk studi klinis yang bertujuan untuk menilai efek obat baru, prosedur atau pencegahan penyakit maka $RR = \text{resiko pada kelompok intervensi} / \text{resiko pada kelompok kontrol}$.
 - $RR=1$, artinya resiko sama pada kedua kelompok dan tidak ada hubungan antara faktor resiko dengan kejadian penyakit.
 - $RR > 1$, artinya resiko penyakit lebih tinggi pada kelompok terpapar (kelompok intervensi) dibandingkan dengan resiko kelompok tidak terpapar (kelompok kontrol).
 - $RR < 1$, artinya resiko lebih rendah pada kelompok terpapar dan menunjukkan bahwa faktor paparan (intervensi) merupakan proteksi
- $p1 = a/m1 = 35/55 = 0,64$
 $p2 = c/m2 = 25/55 = 0,45$
 $\text{Relative Risk (RR)} = p1/p2 = 0,64/0,45 = 1,4$

Kesimpulan:

Orang Hipertensi yang merokok mempunyai resiko 1,4 kali lebih besar terkena penyakit jantung koroner dibanding orang hipertensi yang tidak merokok (pada prospektif studi).

- OR adalah perbandingan kemungkinan peristiwa terjadi dalam satu kelompok dengan kemungkinan hal yang sama terjadi di kelompok lain.
- Dalam uji klinis OR adalah rasio dari odds terjadinya penyakit pada kelompok paparan (kelompok intervensi) dibandingkan dengan odds pada kelompok tidak terpapar (kelompok kontrol).

Faktor Resiko	Penyakit		Total
	Jantung Koroner	Tidak Jantung Koroner	
Merokok	35	20	55
Tidak Merokok	25	30	55
Total	60	50	110

Odds Ratio = rasio p/q

p = rasio penyakit positif

q = rasio penyakit negatif

p = (1- q)

Atau bisa dihitung sebagai berikut:

Odds ratio = (a/c)/(b/d) = ad/bc

a/c = ratio penyakit positif

b/d = ratio penyakit negatif

Odds Ratio (OR) = ad/bc = (35x30)/ (20x25) = 2,1

Kesimpulan:

Orang Hipertensi yang merokok mempunyai resiko 2,1 kali lebih besar terserang penyakit jantung koroner dibanding orang hipertensi yang tidak merokok (pada retrospektif studi).

Latihan dan Tugas

1. Ingin diketahui kecenderungan perempuan untuk memilih tertentu dibandingkan dengan laki-laki, oleh karena itu carilah nilai Odds Ratio dan Relative Risk untuk tabel di bawah ini dan jelaskan artinya!

	Kandidat A	Kandidat B	Total
Laki-Laki	200	406	606
Perempuan	418	418	836
Total	618	824	1442

2. Sebuah studi mengamati 100 perokok dan 100 non-perokok, dan mengikuti mereka untuk pengembangan kanker paru-paru. Penyakitnya adalah kanker paru-paru, paparan merokok, jumlah total masing-masing kelompok adalah 100, dan total semua orang dalam penelitian ini adalah 200. Pada akhir penelitian, mereka menemukan bahwa 30 perokok dan 10 orang non-perokok

mengidap kanker paru-paru. Hitunglah Odds ratio dan relative risk serta jelaskan artinya!

	Lung Cancer	No Lung Cancer
Smoker	30	70
No Smoker	10	90

Materi 9

Analisis Regresi Logistik

Regresi logistik adalah teknik pemodelan yang merupakan pendekatan untuk membuat model prediksi seperti halnya regresi linear (*Ordinary Least Square*). Perbedaannya adalah pada regresi logistik, variabel tidak bebas (*dependent variable*) berupa nilai dikotomi. Skala dikotomi yang dimaksud adalah skala data nominal dengan dua kategori, misalnya: Ya dan Tidak, Baik dan Buruk atau Tinggi dan Rendah. Apabila pada regresi linear (*Ordinary Least Square*) mewajibkan syarat atau asumsi bahwa error varians (residual) terdistribusi secara normal. Sebaliknya, pada regresi ini tidak dibutuhkan asumsi tersebut, sebab pada regresi jenis logistik ini mengikuti distribusi logistik. Dalam regresi logistik dikenal konsep Odds Ratio.

Bentuk umum persamaan regresi logistic:

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \beta_0 + \beta_1 X$$

Dimana: Ln: Logaritma Natural dan exp adalah fungsi exponent (exponen merupakan kebalikan dari logaritma natural. Sedangkan logaritma natural adalah bentuk logaritma dengan nilai konstanta 2,71828182845904 dibulatkan menjadi 2,72).

Model persamaan di atas sulit untuk menginterpretasikan koefisien regresinya. Oleh karena itu maka diperkenalkanlah istilah Odds Ratio (OR) atau yang biasa disingkat $\text{Exp}(\beta)$. $\text{Exp}(\beta)$ merupakan exponent dari koefisien regresi. Jadi misalkan nilai slope dari regresi adalah sebesar 0,80, maka $\text{Exp}(\beta)$ dapat diperkirakan sebagai berikut: $(2,72)^{0,08} = 2,23$

Nilai Odds Ratio

Besarnya nilai $\text{Exp}(\beta)$ dapat diartikan sebagai berikut: Misalnya nilai $\text{Exp}(\beta)$ pengaruh rokok terhadap kanker paru adalah sebesar 2,23, maka disimpulkan bahwa orang yang merokok lebih beresiko untuk mengalami kanker paru dibandingkan dengan orang yang tidak merokok. Interpretasi ini diartikan apabila pengkodean kategori pada tiap variabel sebagai berikut:

1. Variabel bebas adalah Rokok: Kode 0 untuk tidak merokok, kode 1 untuk merokok.
2. Variabel terikat adalah kanker Paru: Kode 0 untuk tidak mengalami kanker paru, kode 1 untuk mengalami kanker paru.

Pseudo R Square

Perbedaan lainnya yaitu pada regresi ini tidak ada nilai “R Square” untuk mengukur besarnya pengaruh simultan beberapa variabel bebas terhadap variabel terikat. Dalam regresi logistik dikenal istilah *Pseudo R Square*, yaitu nilai **R Square Semu** yang maksudnya sama atau identik dengan R Square pada OLS. Jika pada OLS menggunakan uji F Anova untuk mengukur tingkat signifikansi dan seberapa baik model persamaan yang terbentuk, maka pada regresi ini menggunakan Nilai Chi-Square. Perhitungan nilai Chi-Square ini berdasarkan perhitungan Maximum Likelihood.

Materi 10

Analisis Regresi Logistik Multinomial

Regresi Logistik Multinomial adalah analisis regresi linier yang dilakukan bila variabel dependen bersifat nominal dengan lebih dari dua tingkat. Dengan demikian, ini merupakan perpanjangan dari regresi logistik, yang menganalisis tanggungan dikotomis (biner). Karena output SPSS dari analisis agak berbeda dengan hasil regresi logistik, regresi multinomial kadang-kadang digunakan sebagai gantinya.

Regresi multinomial adalah merupakan teknik pemodelan dengan bentuk model persamaan yang link function berbentuk Logit dan responnya berupa data kualitatif nominal dengan jumlah kategorinya lebih dari dua. Sebagai contoh variable responnya berupa wilayah (Jakarta Pusat, Jakarta Barat, Jakarta Selatan), pekerjaan (PNS, Karyawan swasta, dokter, guru dll).

Bentuk Umum Regresi Logistik Multinomial:

$$\text{logit}(y=2) = \log\left(\frac{p(y=2)}{1 - (p=2)}\right) = \beta_0 + \beta_1 \cdot x_{i2} + \beta_2 \cdot x_{i2} + \dots + \beta_p \cdot x_{in} \text{ for } i = 1 \dots n.$$

$$\text{logit}(y=2) = \log\left(\frac{p(y=2)}{1 - (p=2)}\right) = \beta_0 + \beta_1 \cdot x_{i2} + \beta_2 \cdot x_{i2} + \dots + \beta_p \cdot x_{in} \text{ for } i = 1 \dots n.$$

Contoh regresi logistik multinomial

Contoh 1. Pilihan pekerjaan seseorang mungkin dipengaruhi oleh pekerjaan orang tua mereka dan tingkat pendidikan mereka sendiri. Kita bisa mempelajari hubungan pilihan pekerjaan seseorang dengan tingkat pendidikan dan pekerjaan ayah. Pilihan pekerjaan akan menjadi variabel tidak bebas (*dependent variable*) yang terdiri dari kategori pekerjaan.

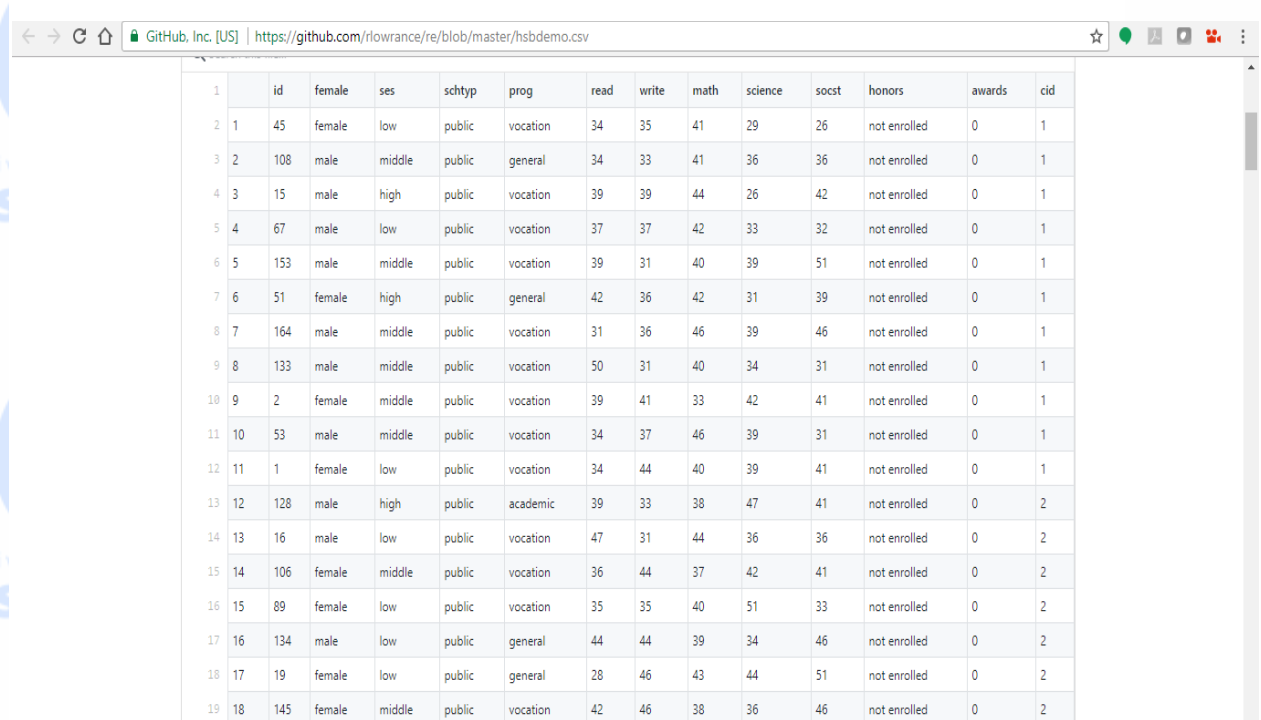
Contoh 2. Memasuki siswa SMA membuat pilihan program antara program umum, program kejuruan dan program akademik. Pilihan mereka mungkin dimodelkan dengan menggunakan nilai tulisan dan status sosial ekonomi mereka.

Deskripsi data:

Kumpulan data berisi variabel pada 200 siswa. Variabel hasil adalah prog, tipe program. Variabel prediktor adalah status sosial ekonomi, ses, variabel kategoris tiga tingkat dan nilai penulisan, tulis, variabel kontinyu. Mari kita mulai dengan mendapatkan beberapa statistik deskriptif tentang variabel minat.

Data asli :

<https://github.com/rlowrance/re/blob/master/hsbdemo.csv>



	id	female	ses	schtyp	prog	read	write	math	science	socst	honors	awards	cid
1	45	female	low	public	vocation	34	35	41	29	26	not enrolled	0	1
2	108	male	middle	public	general	34	33	41	36	36	not enrolled	0	1
3	15	male	high	public	vocation	39	39	44	26	42	not enrolled	0	1
4	67	male	low	public	vocation	37	37	42	33	32	not enrolled	0	1
5	153	male	middle	public	vocation	39	31	40	39	51	not enrolled	0	1
6	51	female	high	public	general	42	36	42	31	39	not enrolled	0	1
7	164	male	middle	public	vocation	31	36	46	39	46	not enrolled	0	1
8	133	male	middle	public	vocation	50	31	40	34	31	not enrolled	0	1
9	2	female	middle	public	vocation	39	41	33	42	41	not enrolled	0	1
10	53	male	middle	public	vocation	34	37	46	39	31	not enrolled	0	1
11	1	female	low	public	vocation	34	44	40	39	41	not enrolled	0	1
12	128	male	high	public	academic	39	33	38	47	41	not enrolled	0	2
13	16	male	low	public	vocation	47	31	44	36	36	not enrolled	0	2
14	106	female	middle	public	vocation	36	44	37	42	41	not enrolled	0	2
15	89	female	low	public	vocation	35	35	40	51	33	not enrolled	0	2
16	134	male	low	public	general	44	44	39	34	46	not enrolled	0	2
17	19	female	low	public	general	28	46	43	44	51	not enrolled	0	2
18	145	female	middle	public	vocation	42	46	38	36	46	not enrolled	0	2

Keterangan setiap variabel:

Id = id peserta

Female (jenis kelamin), dimana: 0 = Male dan 1 = Female

Ses (social economic status), dimana: 1 = low, 2 = middle, 3 = high

Schtyp (school type), dimana: 1 = public, 2 = private

Prog (pilihan program), dimana: 1 = general, 2 = academic, 3 = vocation

Read (reading score)

Write (writing score)

Math (math score)

Science (science score)

Socst (social studies score)

Honors (honors English), dimana 0 = not enrolled dan 1 = enrolled

hsbdemo.sav [DataSet2] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

Visible: 13 of 13 Variables

	id	female	ses	schttyp	prog	read	write	math	science	socst	honors	awards	cid	var	var	var	var	va
1	45	1	1	1	3	34	35	41	29	26	0	0	1					
2	108	0	2	1	1	34	33	41	36	36	0	0	1					
3	15	0	3	1	3	39	39	44	26	42	0	0	1					
4	67	0	1	1	3	37	37	42	33	32	0	0	1					
5	153	0	2	1	3	39	31	40	39	51	0	0	1					
6	51	1	3	1	1	42	36	42	31	39	0	0	1					
7	164	0	2	1	3	31	36	46	39	46	0	0	1					
8	133	0	2	1	3	50	31	40	34	31	0	0	1					
9	2	1	2	1	3	39	41	33	42	41	0	0	1					
10	53	0	2	1	3	34	37	46	39	31	0	0	1					
11	1	1	1	1	3	34	44	40	39	41	0	0	1					
12	128	0	3	1	2	39	33	38	47	41	0	0	2					
13	16	0	1	1	3	47	31	44	36	36	0	0	2					
14	106	1	2	1	3	36	44	37	42	41	0	0	2					
15	89	1	1	1	3	35	35	40	51	33	0	0	2					
16	134	0	1	1	1	44	44	39	51	46	0	0	2					
17	19	1	1	1	1	28	46	43	44	51	0	0	2					
18	145	1	2	1	3	42	46	38	36	46	0	0	2					
19	11	0	2	1	2	34	46	45	39	36	0	0	2					
20	117	0	3	1	3	34	49	39	42	56	0	0	2					
21	109	1	2	1	1	42	39	42	42	41	0	0	2					
22	12	0	2	1	3	37	44	45	39	46	0	0	3					
23	37	1	1	1	3	41	47	40	39	51	0	0	3					

Data View Variable View

IBM SPSS Statistics Processor is ready Unicode:ON

Type here to search

2:45 PM 12/4/2017

hsbdemo.sav [DataSet2] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	id	Numeric	3	0		None	None	5	Right	Scale	Input
2	female	Numeric	1	0		{0, male}...	None	8	Right	Nominal	Input
3	ses	Numeric	1	0		{1, low}...	None	5	Right	Nominal	Input
4	schttyp	Numeric	1	0	type of school	{1, public}...	None	8	Right	Nominal	Input
5	prog	Numeric	1	0	type of program	general...	None	6	Right	Nominal	Input
6	read	Numeric	2	0	reading score	None	None	6	Right	Scale	Input
7	write	Numeric	2	0	writing score	None	None	7	Right	Scale	Input
8	math	Numeric	2	0	math score	None	None	6	Right	Scale	Input
9	science	Numeric	2	0	science score	None	None	9	Right	Scale	Input
10	socst	Numeric	2	0	social studies s...	None	None	7	Right	Scale	Input
11	honors	Numeric	1	0	honors english	{0, not enrol...	None	8	Right	Nominal	Input
12	awards	Numeric	1	0		None	None	8	Right	Nominal	Input
13	cid	Numeric	2	0		None	None	5	Right	Nominal	Input
14											
15											
16											
17											
18											
19											
20											
21											
22											
23											
24											

Data View Variable View

IBM SPSS Statistics Processor is ready Unicode:ON

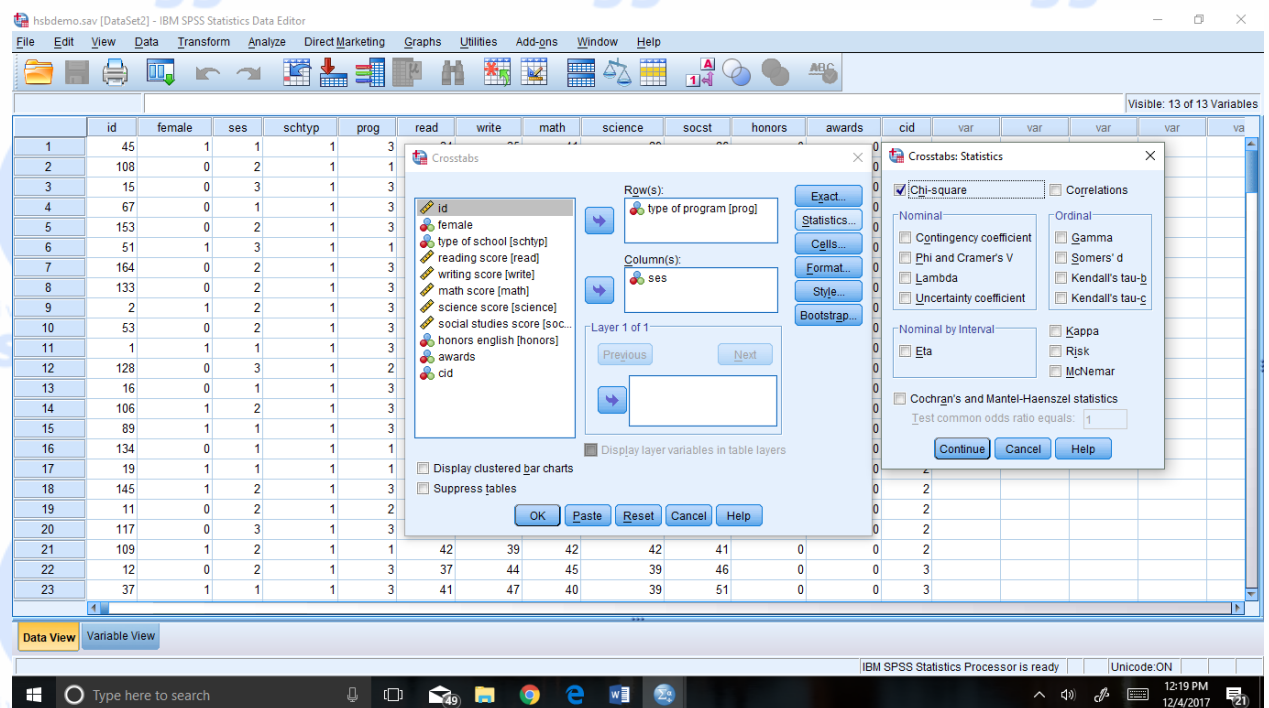
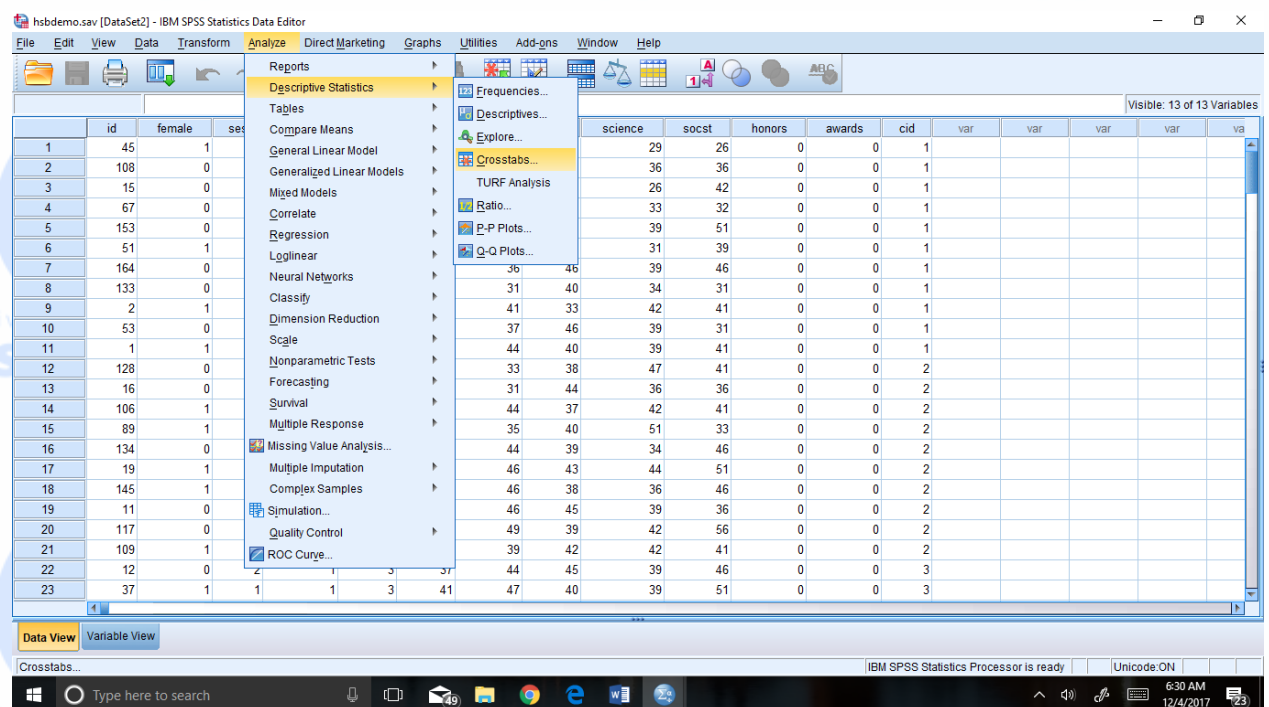
Type here to search

2:45 PM 12/4/2017

Analisis Regresi Logistik Multinomial dengan SPSS:

- Langkah pertama yang dilakukan adalah lakukan analisa deskriptif statistic dengan melihat tabel tabulasinya, hal ini agar mendapatkan gambaran awal bagaimana hubungan antar variabel yang akan diteliti.

Analyze > Descriptive Statistics > Crosstabs



Hasil output crosstab dan chi-square sebagai berikut:

Crosstabs

type of program * ses Crosstabulation

Count

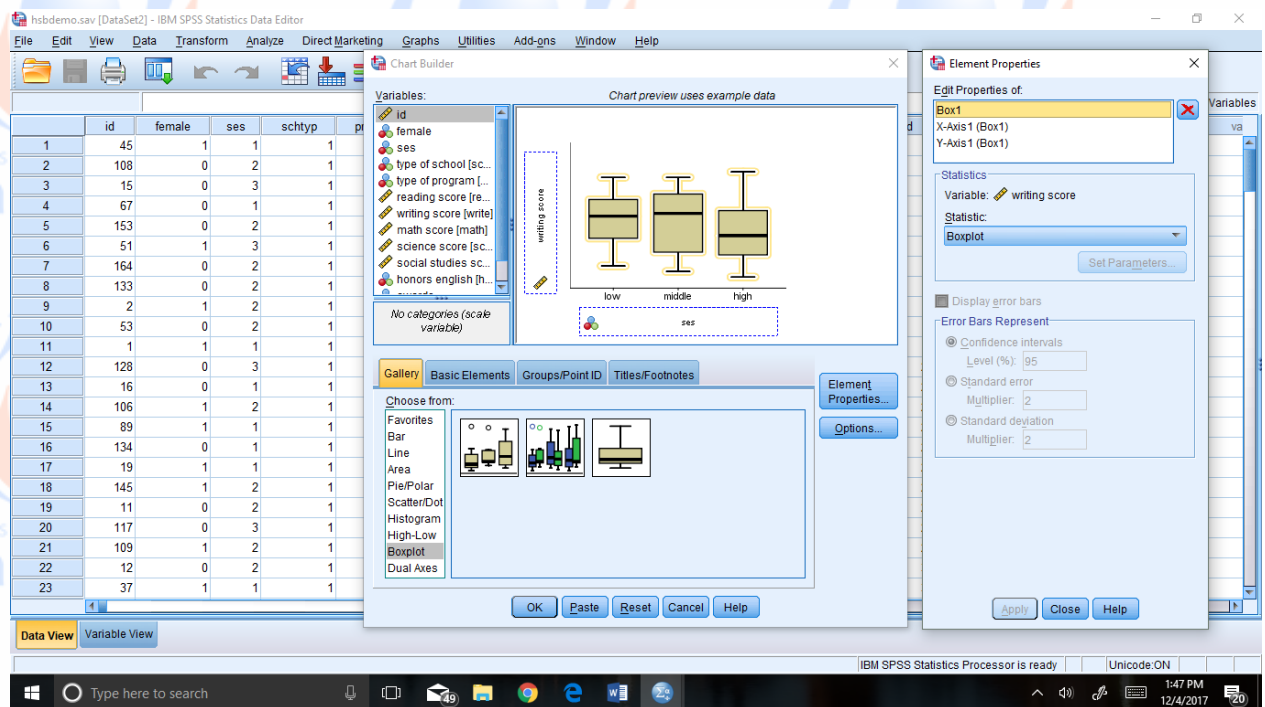
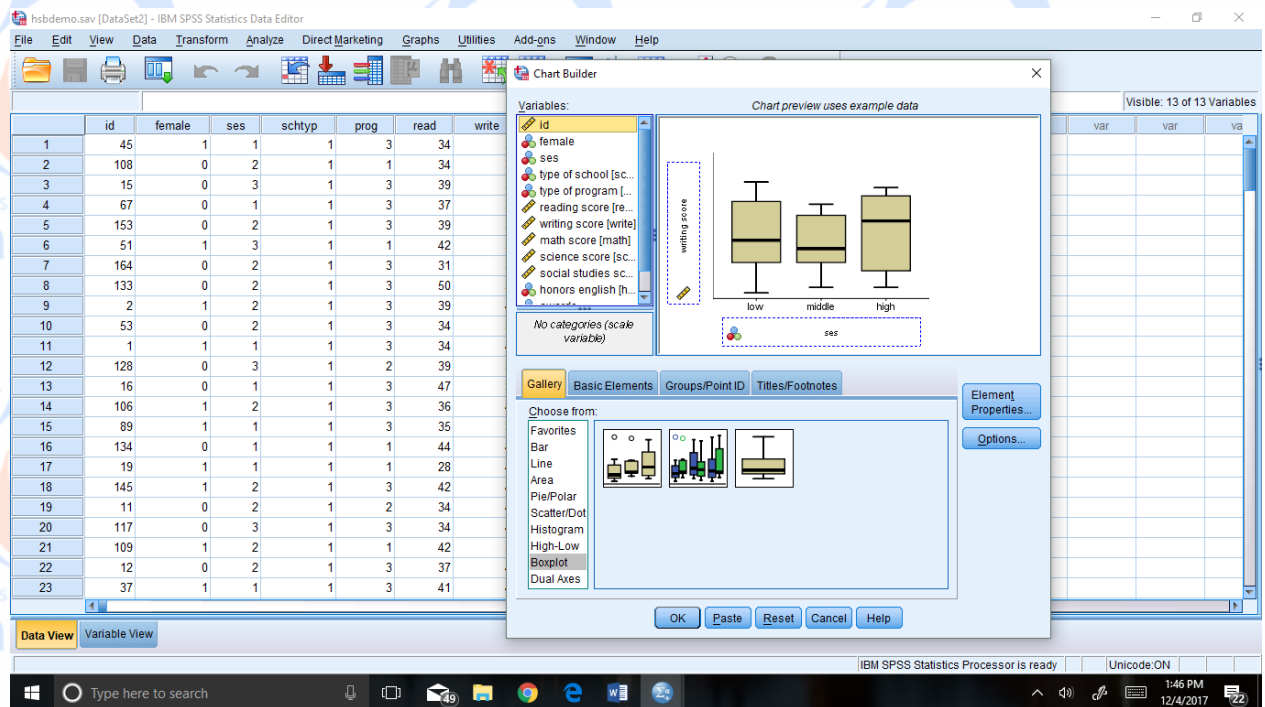
		ses			Total
		low	middle	high	
type of program	general	16	20	9	45
	academic	19	44	42	105
	vocation	12	31	7	50
Total		47	95	58	200

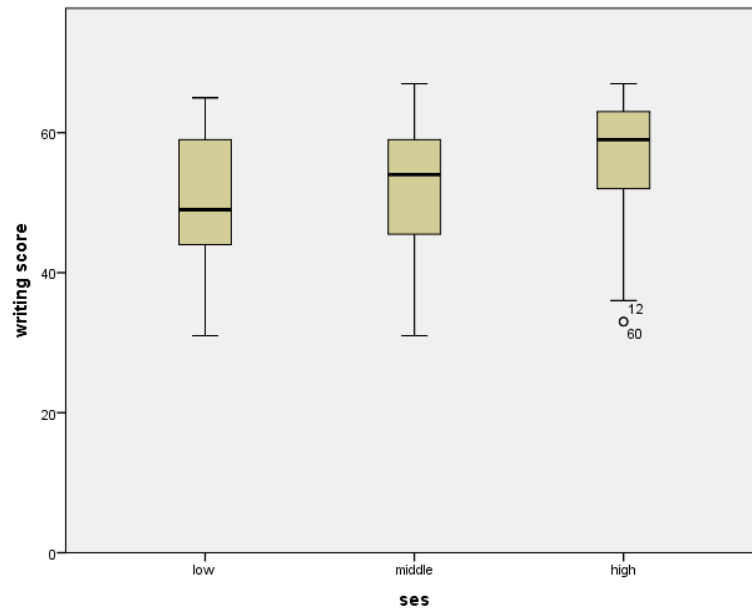
Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	16.604 ^a	4	.002
Likelihood Ratio	16.783	4	.002
Linear-by-Linear Association	.060	1	.807
N of Valid Cases	200		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 10.58.

2. Gambar boxplot:





3. Lakukan analisis regresi logistik multinomial

hsbdemo.sav [DataSet2] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Simulation...
Quality Control
ROC Curve...

Automatic Linear Modeling...
Linear...
Curve Estimation...
Partial Least Squares...
Binary Logistic...
Multinomial Logistic...
Ordinal...
Probit...
Nonlinear...
Weight Estimation...
2-Stage Least Squares...
Optimal Scaling (CATREG)...

	id	female	ses	write	math	science	socst	honors	awards	cid	var	var	var	var	va
1	45	1		35	41	29	26	0	0	1					
2	108	0		33	41	36	36	0	0	1					
3	15	0		39	44	26	42	0	0	1					
4	67	0		37	42	33	32	0	0	1					
5	153	0				39	51	0	0	1					
6	51	1				31	39	0	0	1					
7	164	0				39	46	0	0	1					
8	133	0				34	31	0	0	1					
9	2	1				42	41	0	0	1					
10	53	0				39	31	0	0	1					
11	1	1				39	41	0	0	1					
12	128	0				47	41	0	0	2					
13	16	0				36	36	0	0	2					
14	106	1				42	41	0	0	2					
15	89	1				51	33	0	0	2					
16	134	0				34	46	0	0	2					
17	19	1				44	51	0	0	2					
18	145	1				36	46	0	0	2					
19	11	0				46	45	39	36	0	0	2			
20	117	0				49	39	42	56	0	0	2			
21	109	1				39	42	42	41	0	0	2			
22	12	0				44	45	39	46	0	0	3			
23	37	1	1	1	1	3	41	47	40	39	51	0	0	3	

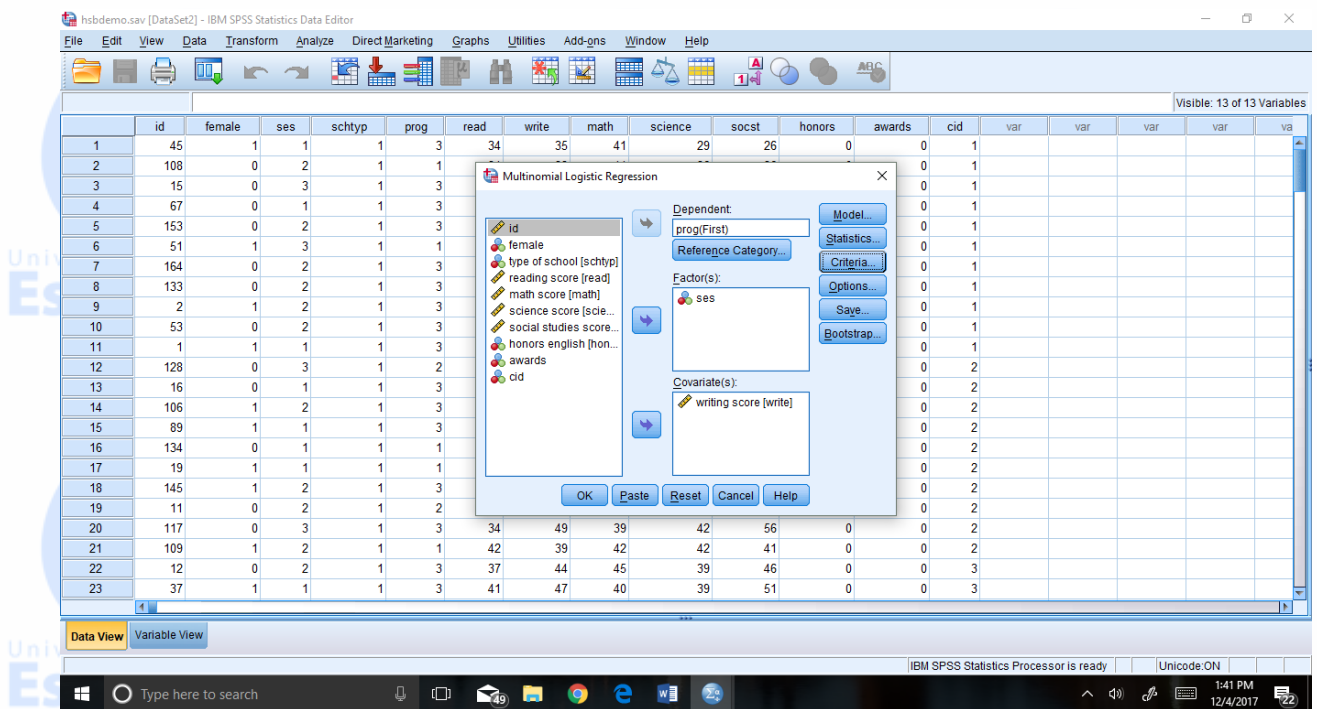
Visible: 13 of 13 Variables

Data View Variable View

Multinomial Logistic...

IBM SPSS Statistics Processor is ready Unicode: ON

1:41 PM 12/4/2017



Nominal Regression

Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	254.986			
Final	206.756	48.230	6	.000

Pseudo R-Square

Cox and Snell	.214
Nagelkerke	.246
McFadden	.118

Likelihood Ratio Tests

Effect	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	206.756 ^a	.000	0	.
write	238.203	31.447	2	.000
ses	217.815	11.058	4	.026

Parameter Estimates

		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
academic	Intercept	-1.689	1.227	1.896	1	.169			
	write	.058	.021	7.320	1	.007	1.060	1.016	1.105
	[ses=1]	-1.163	.514	5.114	1	.024	.313	.114	.856
	[ses=2]	-.630	.465	1.833	1	.176	.533	.214	1.326
	[ses=3]	0 ^b	.	.	0
vocational	Intercept	2.546	1.278	3.967	1	.046			
	write	-.056	.023	5.694	1	.017	.946	.904	.990
	[ses=1]	-.180	.648	.077	1	.781	.835	.234	2.977
	[ses=2]	.645	.592	1.184	1	.276	1.905	.597	6.082
	[ses=3]	0 ^b	.	.	0

a. The reference category is: general.

b. This parameter is set to zero because it is redundant.

Rasio kemungkinan chi-square 48,23 dengan nilai $p < 0,0001$ memberi tahu kita bahwa model kita secara keseluruhan sangat jauh lebih baik daripada model kosong (yaitu model tanpa prediktor)

Output di atas memiliki dua bagian, diberi label dengan kategori variabel hasil prog. Mereka sesuai dengan dua persamaan di bawah ini:

$$\ln \left(\frac{P(\text{prog} = \text{general})}{P(\text{prog} = \text{academic})} \right) = b_{10} + b_{11}(\text{ses} = 1) + b_{12}(\text{ses} = 2) + b_{13}\text{write}$$

$$\ln \left(\frac{P(\text{prog} = \text{vocation})}{P(\text{prog} = \text{academic})} \right) = b_{20} + b_{21}(\text{ses} = 1) + b_{22}(\text{ses} = 2) + b_{23}\text{write}$$

dimana b adalah koefisien regresi.

Peningkatan satu unit dalam variabel tulis dikaitkan dengan penurunan 0,00 dalam log relatif kemungkinan berada dalam program umum versus program akademik.

Peningkatan satu unit dalam variabel tulis dikaitkan dengan penurunan nilai log relatif di dalam program panggilan versus program akademik .1136.

Kemungkinan log relatif berada dalam program umum versus dalam program akademik akan meningkat sebesar 1,163 jika berpindah dari tingkat tertinggi ses (ses = 3) ke tingkat terendah ses (ses = 1).

Rasio probabilitas untuk memilih satu kategori hasil karena probabilitas memilih kategori baseline sering disebut sebagai risiko relatif (dan kadang kala disebut sebagai peluang karena kita baru saja menggunakan parameter regresi di atas).

Dengan demikian, eksponensial persamaan linier di atas menghasilkan risiko relatif.

Koefisien regresi mewakili perubahan risiko relatif log (log odds) per unit perubahan pada prediktor. Koefisien regresi eksponensial akan menghasilkan rasio risiko relatif. SPSS mencakup rasio risiko relatif dalam output, di bawah kolom "Exp (B)".

Rasio risiko relatif untuk kenaikan satu unit dalam variabel tulis adalah 0,9437 (exp (- 0579284) dari keluaran perintah nomreg di atas) karena mengikuti program umum versus program akademik.

Rasio risiko relatif yang beralih dari $ses = 3$ menjadi 1 adalah 3.199 karena mengikuti program umum versus program akademik. Dengan kata lain, risiko yang diharapkan untuk bertahan dalam program umum lebih tinggi untuk subjek yang rendah ses. Pengujian untuk keseluruhan efek ses dan write dikeluarkan oleh perintah nomreg. Di bawah ini kita melihat bahwa dampaknya signifikan secara statistik.

Tugas: (dilakukan secara kelompok)

Kasusnya masih sama dengan latihan yaitu tentang pemilihan program siswa SMA apakah dipengaruhi dari status sosial ekonominya dan kemampuan menulis mereka. Adapaun jenis program yang akan dipilih terdiri dari 3 kategori yaitu: general (umum), akademik (academic) dan kejuruan (vocation).

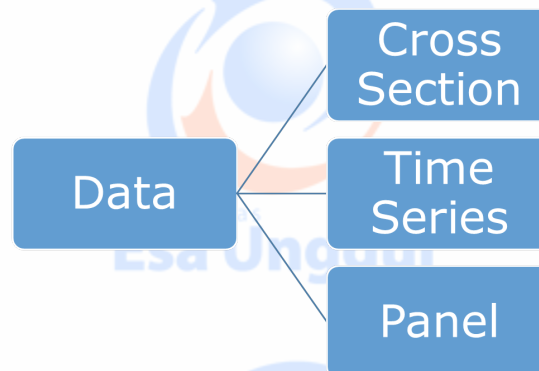
Tambahkan 1 faktor lain (bebas memilih: misal tipe sekolah (variabel schtype), reading score dll) yang menurut anda ada pengaruhnya terhadap pemilihan program studi tersebut! Data bisa di download. Data bernama hsbdemo.sav ([data latihan regresi logistik multinomial](#))



Materi 11

Analisis Deret Waktu

Pada bagian ini mahasiswa akan mempelajari tentang data deret waktu serta mempelajari komponen apa saja yang ada di dalam deret waktu. Secara garis besar, berdasarkan waktunya data dapat dibagi menjadi:



1. Data *Time Series* (runtun waktu/historis)

Data *time series* adalah data yang terdiri dari satu objek namun teridiri dari beberapa waktu periode, seperti harian, bulanan, triwulanan, dan tahunan.

Misalnya: data jumlah pasien rawat inap di sebuah rumah sakit tahun 2010 sampai 2017 (objeknya hanya satu: jumlah pasien rawat inap namun disajikan dalam beberapa periode: dari tahun 2010-2017 secara triwulan).

2. Data Silang (*cross section*)

Data Silang (*cross section*) adalah data yang terdiri dari satu objek yang dikumpulkan pada suatu waktu (satu waktu saja, tidak seperti data time series yang terdiri dari beberapa periode waktu).

3. Data Panel (*pooled data*)

Data Panel (*pooled data*) adalah data gabungan antara data *time series* dengan data *cross section*. Dikatakan data gabungan karena data ini terdiri atas beberapa objek/sub objek dalam beberapa periode waktu.

Eksplorasi Data

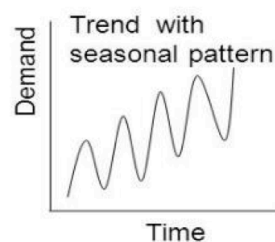
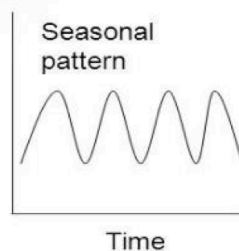
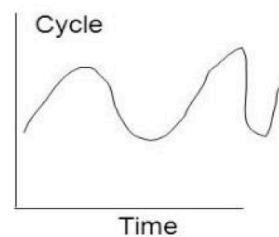
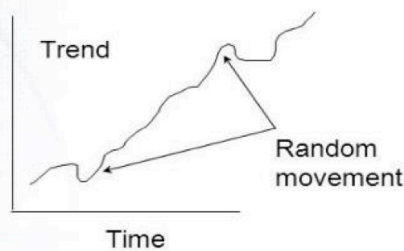
Ketika bekerja dengan data time series atau deret waktu, hal pertama yang penting dilakukan adalah membuat plot data agar mengetahui gambaran atau pola datanya seperti apa, baru kemudian bisa melakukan analisis lebih lanjut.

Pola data deret waktu terdapat empat komponen utama, yaitu:

7. Trend, yaitu komponen jangka panjang yang mendasari pertumbuhan (atau penurunan) suatu data runtut waktu. Merupakan pergerakan data sedikit demi sedikit meningkat atau menurun.
8. Siklus (cyclical), yaitu suatu pola dalam data yang terjadi setiap beberapa tahun. fluktuasi atau siklus dari data runtut waktu akibat perubahan kondisi ekonomi
9. Musiman (seasonal), yaitu pola data yang berulang pada kurun waktu tertentu. fluktuasi musiman yang sering dijumpai pada data kuartalan, bulanan atau mingguan.
10. Acak (random), yaitu pola acak yang disebabkan oleh peristiwa yang tidak bisa diprediksi atau tidak beraturan.



Time Series Components



Kegunaan Analisis Deret Waktu:

1. Untuk Peramalan atau *forecasting*: adalah suatu proses untuk memprediksi kejadian masa depan. Hal ini juga menjadi dasar untuk mengambil keputusan di bidang: *Production, Inventory, Personnel and Facilities*
2. Untuk Pemodelan

Konsep Dasar Deret Waktu:

- **Proses Deterministik** : jika dari pengalaman yang lalu keadaan yang akan datang dari suatu barisan kejadian dapat diramalkan secara pasti, maka barisan kejadian itu dinamakan deterministik
- **Proses Stokastik**: jika pengalaman yang lalu hanya dapat menyajikan struktur peluang keadaan yang akan datang, maka barisan kejadian yang demikian disebut stokastik. Hull, 1989 → setiap nilai yang berubah terhadap waktu dengan cara yang tidak tertentu (dalam ketidakpastian) dikatakan mengikuti proses stokastik.

Latihan:

1. Jelaskan perbedaan data cross sectional dan data deret waktu!
2. Berikan contoh data deret waktu!
3. Sebutkan dan jelaskan komponen dalam analisis deret waktu!

Materi 12

Analisis Deret Waktu

Pada tahap ini mahasiswa akan mempelajari beberapa model atau metode dalam analisis deret waktu. Sebenarnya ada banyak metode dalam analisis deret waktu tergantung dari komponen yang terdapat dalam runtun waktu tersebut. Teknik peramalan data deret waktu yang ada diantaranya

- AR (Auto Regressive) Model
- MA (Moving Average) Model
- ARIMA (Autoregressive Integrated Moving Average) Model

Metode ini biasanya dipakai untuk prediksi atau peramalan baik jangka pendek maupun jangka panjang. Disesuaikan dengan kebutuhan dan struktur data atau komponen time seriesnya. Tetapi dalam modul ini untuk tahap awal akan dikenalkan metode yang paling sederhana yaitu: Metode Trend dan Metode Rata-Rata Bergerak (*Moving Average method*).

■ Linear Trend (Least Square Method)

$$Y_t = a + bx$$

Untuk pengerjaannya dapat dilakukan dengan 2 cara yang hasilnya semuanya sama.

Cara 1: membuat penjumlahan dari x sama dengan nol

Cara 2: membuat nilai X berurut dari 1 sampai dengan n (yang terakhir)

Cara 1 Metode Least Square:

Tahun	X	Y	XY	X ²
1992	-7	10.164,9	-71.154,3	49
1992	-5	11.169,2	-55.846,0	25
1994	-3	12.054,6	-36.163,8	9
1995	-1	12.325,4	-12.325,4	1
1996	1	12.842,2	12.842,2	1
1997	3	13.511,5	40.534,5	9
1998	5	14.180,8	70.904,0	25
1999	7	14.850,1	103.950,7	49
Jumlah (Σ)		$\Sigma y = 101.098,7$	$\Sigma xy = 52.741,9$	$\Sigma x^2 = 168$

$$a = y = 101.098,7 / 8 = \mathbf{12.637,34}, \quad \text{dimana } n = 8$$

$$b = (\sum X Y) / (\sum X^2) = 52.741,9 / 168 = \mathbf{313,94}$$

Maka persamaan trend linier secara *Least Square method* adalah

$$Y = a + bx,$$

$$\text{dimana } a = 12.637,34 \quad \text{dan} \quad b = 313,94$$

$$Y = \mathbf{12.637,34 + 313,94 X}$$

Untuk meramalkan nilai y tahun 2000, nilai x menjadi = 9, masukan kedalam persamaan diatas, sehingga

$$Y = \mathbf{12.637,34 + 313,94 (9) = 15.462,8}$$

Jadi ramalan Pendapatan tahun 2000 = Rp. 15.462,8 milyar

Cara 2 Metode Least Square

Tahun	X	Y	XY	X ²
1992	1	10.164,9	10.164,9	1
1992	2	11.169,2	22.338,0	4
1994	3	12.054,6	36.163,8	9
1995	4	12.325,4	49.301,6	16
1996	5	12.842,2	64.211,0	25
1997	6	13.511,5	81.069,0	36
1998	7	14.180,8	99.265,6	49
1999	8	14.850,1	118.800,8	64
(Σ)	Σx = 36	Σy = 101.098,7	Σxy = 481.315,1	Σx ² = 204

Mencari b dan a

$$b = \frac{n (\sum X Y) - (\sum X)(\sum Y)}{n (\sum X^2) - (\sum X)^2} = \frac{3.850.520,8 - 3.639.553,2}{1.632 - 1.296} = \mathbf{627,879}$$

$$a = Y - bX = 12.637,34 - 627,88 (4,5) = \mathbf{9.811,88}$$

$$\text{Persamaan menjadi } Y = \mathbf{9.811,88 + 627,879 X}$$

Untuk meramalkan nilai y tahun 2000, nilai x menjadi = 9, masukan ke dalam persamaan di atas, sehingga

$$y = \mathbf{9.811,88 + 627,879 (9) = 15.462,8}$$

Jadi ramalan Pendapatan tahun 2000 = Rp. 15.462,8 milyar (= dengan cara 1)

Metode Rata-rata Bergerak (Moving Average Methods)

Dalam metode ini prinsip yang digunakan adalah dengan membuat rata-rata berdasarkan berapa besarnya tahun yang akan dihitung. Misalkan rata-rata bergerak tiga tahunan artinya kita menghitung rata-rata tidak tahun pertama, kemudian diulang langkah di atas untuk tahun-tahun berikutnya dengan bergerak setiap 3 tahunan. Begitu juga jika dihitung rata-rata bergerak 5 tahunan, maka dihitung rata-rata 5 tahun pertama dan selanjutnya bergerak ke tahun berikutnya.

Tahun	Y (= jutaan rupiah)	Rata-rata bergerak 4 tahun	Rata-rata bergerak 5 tahun
(1)	(2)	(3)	(4)
1989	50,0		
1990	36,5		
1991	43,0	43,5*	42,6**
1992	44,5	40,7	40,2
1993	38,9	41,1	39,4
1994	38,1	38,5	39,6
1995	32,6	37,1	38,0
1996	38,7	37,8	38,4
1997	41,7	38,5	37,6
1998	41,1	38,8	
1999	33,8		

Materi 13

Ukuran Akurasi Model Deret Waktu

- Mahasiswa mampu memilih model terbaik dengan melihat ukuran akurasi model
- Mempelajari berbagai ukuran keakuratan model, yaitu: MAD, MAPE dan MSE
- Menghitung nilai statistik MAPE, MAD dan MSD untuk dibandingkan, sehingga bisa dibuat untuk pengambilan keputusan metode mana yang paling baik dengan melihat beda antara nilai ramalan dengan nilai aktualnya. Dengan kata lain pilih yang memberikan error terkecil.
- Untuk ketiga ukuran di atas, nilai terkecil yang memberikan indikasi sebagai model terbaik karena memiliki eror yang paling kecil.

Proses Pembuatan Model Time Series

Secara garis besar tahapan pembentukan model deret waktu melalui langkah-langkah berikut ini:

1. Perhatikan Data
2. Gunakan berbagai metode untuk prediksi atau ramalan
3. Evaluasi teknik yang sudah dipilih
4. Pilih model terbaiknya

Observations from the scatter Plot	Techniques to try	Ways to evaluate
Data is reasonably stationary (no trend or seasonality)	Heuristics - Averaging methods <ul style="list-style-type: none"> • Naive • Moving Averages • Simple Exponential Smoothing 	<ul style="list-style-type: none"> • MAD • MAPE • Standard Error • BIAS
Data shows a consistent trend	Regression <ul style="list-style-type: none"> • Linear • Non-linear Regressions (not covered in this course) 	<ul style="list-style-type: none"> • MAD • MAPE • Standard Error • BIAS • R-Squared
Data shows both a trend and a seasonal pattern	Classical decomposition <ul style="list-style-type: none"> • Find Seasonal Index • Use regression analyses to find the trend component 	<ul style="list-style-type: none"> • MAD • MAPE • Standard Error • BIAS • R-Squared

- BIAS - The arithmetic mean of the errors

$$\text{BIAS} = \frac{\sum (\text{Actual} - \text{Forecast})}{n} = \frac{\sum \text{Error}}{n}$$

- n is the number of forecast errors
- Excel: =AVERAGE(error range)

- Mean Absolute Deviation - MAD

$$\text{MAD} = \frac{\sum |\text{Actual} - \text{Forecast}|}{n} = \frac{\sum |\text{Error}|}{n}$$

- No direct Excel function to calculate MAD

Evaluasi Model Deret Waktu:

- Mean Square Error - MSE
 - Excel: =SUMSQ(error range)/COUNT(error range)
 - Standard error is square root of MSE
- Mean Absolute Percentage Error - MAPE
- R^2 - only for curve fitting model such as regression
- In general, the lower the error measure (BIAS, MAD, MSE) or the higher the R^2 , the better the forecasting model

Latihan dan Tugas:

Kerjakan dengan kelompok masing-masing:

1. Cari contoh kasus yang berhubungan dengan area kerja sebagai rekam medis, tenaga kesehatan ataupun fisioterapi, yang menggambarkan data dikumpulkan menurut runtun atau urutan waktu, misal total kunjungan pasien dalam beberapa bulan terakhir di sebuah rumah sakit X, jumlah pasien yang menderita penyakit X dari tahun ke -x sampai tahun ke- Y, dll. Satu soal dikerjakan dengan metode moving average atau metode trend (pilih salah satu metode)!
2. Menurut anda bagaimana memilih model terbaik untuk kasus data deret waktu!



Materi 14

Final Student's Discussion Project

Mahasiswa mampu menganalisa suatu kasus untuk dipecahkan dengan teknik pemodelan statistik dan mempresentasikannya dalam bentuk tulisan dan lisan.

Diskusi Projek Akhir (Final Student Project Discussion)

- Akan diberikan studi kasus
- Mahasiswa diminta untuk melakukan analisa dan membuat tulisan atau makalah serta mempresentasikan di depan kelas.
- Cakupan materi dari pertemuan 8 - 13
- Diskusi Projek Akhir (Final Student Project Discussion)
- Akan diberikan studi kasus
- Mahasiswa diminta untuk melakukan analisa dan membuat tulisan atau makalah serta mempresentasikan di depan kelas.
- Cakupan materi dari pertemuan 8 – 13

KASUS

1. Kasus 1: Dipilih secara random topiknya (OR, RR, dan Multinomial logistic Regression), sumber data dapat diperoleh dari internet dengan mencantumkan sumbernya.
2. Kasus 2: Pemecahan kasus peramalan deret waktu dengan menggunakan metode moving average.
Berikut ini adalah data rata-rata kunjungan harian pasien yang masuk di Unit Gawat Darurat (UGD) sebuah rumah sakit. Data dikumpulkan dari minggu pertama bulan desember tahun 2016 sampai data minggu ke-4 bulan April 2017.
 - a. Prediksikan atau ramalkan rata-rata kunjungan harian pasien UGD dari bulan Desember minggu ke-3 2016 sampai dengan April minggu ke-2 tahun 2017 dengan metode 3-rata-rata bergerak (Moving Average-3 rata-rata)!
 - b. Buatlah plot data actual dan data ramalan dengan menggunakan moving average -3 (3 rata-rata) !

Kasus untuk Metode Moving Average:

Week	Average Daily Admissions
December Week 1	11
December Week 2	14
December Week 3	17
December Week 4	15
Januari Week 1	12
Januari Week 2	11
Januari Week 3	9
Januari Week 4	9
February Week 1	12
February Week 2	8
February Week 3	13
February Week 4	11
March Week 1	15
March Week 2	17
March Week 3	14
March Week 4	19

Daftar Pustaka

1. Kurtner, MH, Nachtsheim CJ, Neter J, Li W. *Applied Linear Statistical Models*. Fifth Edition. McGraw-Hill. 2005.
2. Agresti A. *Categorical Data Analysis*. John Wiley & Sons. 1990.
3. Rosner, B. *Fundamentals of Biostatistics*. Brooks Cole. 2015.
4. Carol, E. *Statistical Applications for Health Information Management*. Aspen Publishers. 2000.
5. Oscan, Y A. *Quantitative Methods in Health Care Management: Techniques and Applications*. John Wiley & Sons. 2005.
6. <http://www.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/Chapter%20%206%20Data%20Sets.html>
7. <https://statistics.laerd.com/spss-tutorials/multinomial-logistic-regression-using-spss-statistics.php>
8. <https://mysite.science.uottawa.ca/rkulik/mat3378/mat3378-textbook.pdf>