REGRASI LINEAR MUDAH: KAJIAN KES JUMLAH KEMALANGAN JALAN RAYA DI MALAYSIA BARAT

Syed Othmawi b Abd Rahman,
P. Madya Jamilin b Jais,
Fakulti Sains Komputer dan Sistem Maklumat,
Universiti Teknologi Malaysia,
Kuala Lumpur

ABSTRAK

Regrasi linear mudah merupakan sebuah model linear yang menghubungkan di antara dua pembolehubah. Kertas kerja ini merupakan kertas kerja asas bagi membincangkan model tersebut. Antara perkara-perkara yang dibincangkan adalah terdiri dari model regrasi mudah yang umum, andaian-andaian yang diperlukan sewaktu menggunakan model tersebut, pengiraan bagi mendapatkan pembolehubah persamaan regrasi, penilaian kebagusan model yang telah dihasilkan dan bagaimana kita boleh menggunakan model tersebut di dalam proses ramalan dan anggaran. Kajian kes kadar kemalangan jalaraya berbanding dengan bilangan kenderaan di atas jalan raya di Malaysia Barat telah digunakan di dalam perbincangan ini. Di akhir sekali, pakej SAS/STAT digunakan bagi melakukan analisis yang sama.

Katakunci: Regrasi linear mudah, model, SAS/STAT, persamaan linear dan kadar kemalangan jalan raya di Malaysia Barat

ABSTRACT

Simple linear regression is a model that incorporates the two variables in a linear relationship. This paper tried to introduce the basic discussions of the state model. The discussion included a general linear regression model, assumptions, calculated parameters and how the model was used in the prediction and assumption. The SAS/STAT package was used in the analysis for the problem solving.

Keywords: Simple linear regression, model, SAS/STAT, linear equations and the rate of road accident in West Malaysia.

1.0 PENDAHULUAN

Di dalam menganalisa data, kadangkala kita ingin mengetahui hubungan di antara satu set data berbanding dengan set yang lain. Contohnya apakah hubungan antara kos iklan yang telah dibelanjakan berbanding dengan jumlah barang yang dibeli oleh pelanggan. Contoh lain, di dalam ujikaji makmal, kita mungkin perlukan hubungan di antara jumlah logam yang dicampur dengan kadar kekuatan bahan yang dihasilkan. Salah satu dari kaedah yang biasa digunakan untuk mendapat nilai hubungan ini adalah analisis regrasi. Menggunakan model tersebut nilai suatu pembolehubah itu dapat diramalkan apabila nilai pembolehubah-pembolehubah lain diketahui. Kaedah ini telah mula digunakan oleh ahli sains British, Sir Fracis Galton (1922-1911) di dalam penyelidikannya ke atas buah-buahan dan manusia.

1.1 MODEL ANALISIS REGRASI YANG MUDAH

Di sini kajian ditumpukan terhadap hubungan di antara bilangan kemalangan jalan raya dengan jumlah kenderaan di atas jalan raya di Malaysia Barat. Awal-awal lagi, kita mungkin telah membuat kesimpulan bahawa apabila bilangan kenderaan di atas jalan raya meningkat, kadar kemalangan juga turut bertambah atau secara lebih mudah lagi kita mungkin beranggapan bahawa hubungan kedua-dua pembolehubah ini linear berkadar terus. Tetapi, sejauh manakah kenyataan ini benar. Bagaimana hubungan kedua-dua pembolehubah ini boleh ditakrifkan.

1.1.1 Persamaan Regrasi

Dalam model regrasi linear mudah ia hanya menarifkan hubungan linear di antara dua pembolehubah, X dan Y, iaitu kita mendapatkan suatu garisan lurus yang menarifkan hubungan di antara kedua-dua pembolehubah. Dengan ini kita boleh tarifkan hubungan tersebut di dalam persamaan berikut:

$$y_i = \alpha + \beta x_i + e_i \qquad (1)$$

Di mana y_i dan x_i adalah nilai-nilai bagi pembolehubah Y dan X yang berkaitan manakala α dan β adalah parameter-parameter ***konstan regrasi***. Nilai e_i pula merupakan pembolehubah rawak dengan purata 0 dan varian σ^2. Nilai ini merupakan ***kesalahan ramalan*** iaitu perbezaan antara nilai y_i dengan nilai ramalan yang dihasilkan oleh persamaan regrasi. Oleh kerana model (1) terdiri dari satu pembolehubah tak bersandar X sahaja, jadi ia dikenali dengan ***model regrasi mudah***.

Tujuan asas analisa regrasi adalah untuk menganggarkan nilai-nilai α dan β. Setelah nilai-nilai ini didapati kita boleh membentuk garisan regrasi Y ke atas X. Dengan ini hubungan di antara pembolehubah tak bersandar X dan pembolehubah bersandar Y dapat diketahui.

1.1.2 Andaian-andaian Model Regrasi Linear Mudah

Pembolehubah X dipanggil pembolehubah tak bersandar manakala pembolehubah Y dipanggil pembolehubah bersandar. Pembolehubah X dipanggil pembolehubah tak bersandar kerana ia boleh mengambil sebarang nilai manakala pembolehubah Y dipanggil pembolehubah bersandar kerana nilainya bergantung terus kepada nilai X. Oleh kerana model yang dibincangkan hanya model regrasi linear mudah, hanya satu pembolehubah X sahaja diperlukan. Walaubagaimana pun sebelum kita dapat menggunakan model analisis regrasi, kita memerlukan andaian-andaian berikut[1]:

1. Nilai-nilai bagi pembolehubah tak bersandar, X boleh ditetapkan ataupun dipilih secara rawak. Ini bermakna kita boleh memilih nilai X terlebih dahulu. Dengan ini sewaktu kita mengumpul data kita kawal nilai X. Atau, kita mendapatkan nilai X tanpa mengenakan sebarang sekatan. Nilai X seperti ini dikatakan rawak. Apabila nilai X yang digunakan itu tidak rawak, model regrasi tersebut dikatakan ***model regrasi klasik***.

2. Nilai X diukur tanpa sebarang kesalahan.

3. Untuk setiap nilai X terdapat subpopulasi nilai-nilai Y. Supaya anggaran dan pengujian hipotesis menjadi sah, subpopulasi-subpopulasi ini mestilah bertaburan normal.

4. Varian subpopulasi-subpopulasi Y adalah sama.

5. Purata subpopulasi-subpopulasi Y semuanya berada di atas garisan lurus. Andaian ini dipanggil andaian ***kelinearan***. Secara simbol ia boleh ditulis seperti berikut:

$$\mu_{y|x} = \alpha + \beta x_i \quad (2)$$

Di mana $\mu_{y|x}$ adalah purata subpopulasi nilai-nilai Y yang diandaikan wujud untuk x_i.

6. Nilai-nilai Y adalah bebas di antara satu sama lain. Ini bererti nilai Y yang didapati bagi setiap nilai X tidak akan bergantung kepada nilai-nilai Y yang didapati oleh nilai X yang lain.

Daripada persamaan (1), kita boleh tulis e_i sebagai:

$$e_i = y_i - (\alpha + \beta x_i) \quad (3)$$

e_i menunjukkan jumlah y_i yang tersisih dari purata subpopulasi nilai-nilai Y. Subpopulasi nilai-nilai Y dianggap bertaburan normal dengan varian yang sama. Oleh itu nilai e_i untuk setiap subpopulasi juga bertaburan normal dengan varian σ^2, varian yang biasa bagi nilai-nilai subpopulasi Y. e_i diandaikan bebas dan taburan mempunyai purata 0.

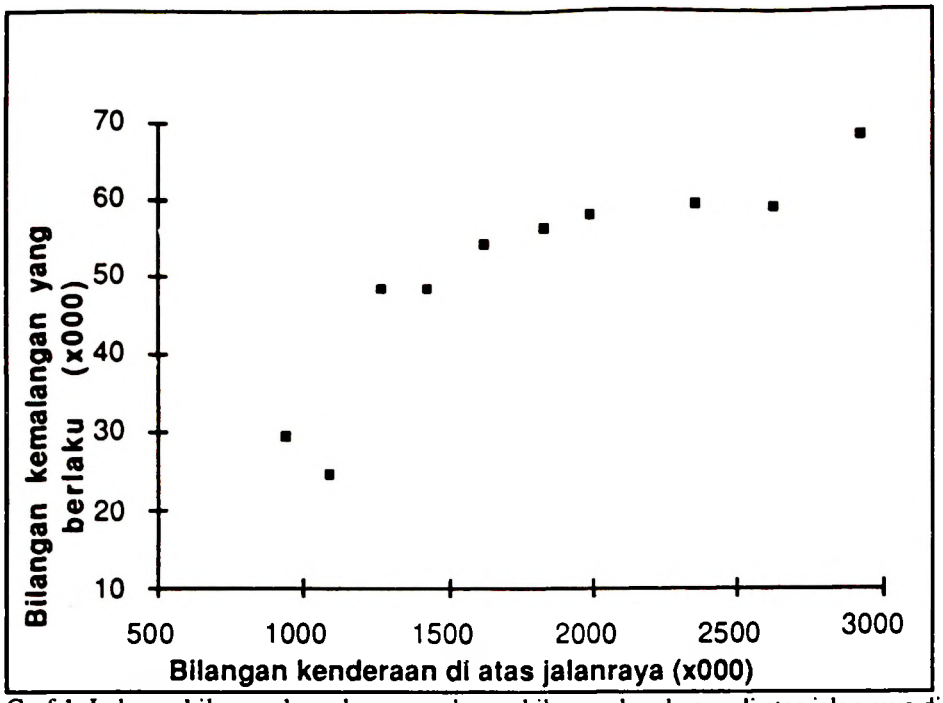
2.0 KAJIAN KES: HUBUNGAN ANTARA BILANGAN KEMALANGAN YANG BERLAKU DENGAN BILANGAN KENDERAAN DI ATAS JALANRAYA

Untuk mengkaji hubungan di antara kadar kemalangan yang berlaku di Malaysia Barat dengan bilangan kenderaan di atas jalan raya, data bagi faktor tersebut dari tahun 1973-1982 akan digunakan sebagai sampel. Oleh itu hanya dua pembolehubah yang terlibat di dalam model tersebut. Bilangan kenderaan di atas jalan raya merupakan pembolehubah tak bersandar (X) manakala bilangan kemalangan yang berlaku merupakan pembolehubah bersandar (Y). Data-data bagi kajian ini ditunjukkan dalam jadual di bawah:

Tahun	1973	1974	1975	1976	1977	1978
Kenderaan di atas jalan raya	939,951	1,090,279	1,267,119	1,429,845	1,621,271	1,829,958
Bilangan kemalangan yang berlaku	29,286	24,581	48,233	48,291	54,222	56,021
Tahun	1979	1980	1981	1982		
Kenderaan di atas jalan raya	1,989,391	2,357,386	2,631,948	2,930,101		
Bilangan kemalangan yang berlaku	57,931	59,084	58,768	68,330		

Jadual 1: Data bilangan kenderaan di atas jalan raya dan bilangan kemalangan yang berlaku di Malaysia Barat.
(Sumber: Statistical Report Road Accidents Malaysia, Royal Malaysia Police, 1982)

Langkah pertama dalam mengkaji hubungan di antara X dan Y adalah dengan melakarkan graf bagi menunjukkan hubungan secara kasar di antara kedua-dua pembolehubah.



Graf 1: Lakaran bilangan kemalangan melawan bilangan kenderaan di atas jalan raya di Malaysia Barat dari tahun 1973 hingga 1982

Daripada lakaran di atas kita mengesyaki terdapat hubungan linear di antara kedua-dua pembolehubah. Dengan ini kita boleh membuat kesimpulan bahawa kadar kemalangan jalan raya di Malaysia akan bertambah apabila bilangan kenderaan di atas jalan raya bertambah. Untuk mendapat garisan hubungan ini, kita boleh melakarkan terus garisan di atas graf tetapi adakah garisan ini merupakan yang terbaik untuk mewakili data X dan Y. Suatu kaedah yang biasa digunakan untuk mendapatkan garisan terbaik adalah ***kaedah kuasadua terkecil.***

2.1 KAEDAH KUASADUA TERKECIL

Menggunakan kaedah ini persamaan garislurus tersebut ditulis seperti berikut:

$$y = a + bx \qquad (4)$$

a adalah titik dimana garisan memotong paksi Y dan b adalah jumlah pertambahan nilai Y disebabkan oleh pertambahan satu unit nilai X. Kita memanggil ***a*** sebagai ***nilai pemotongan*** dan ***b*** adalah ***nilai kecerunan garisan***. Menggunakan kedua-dua nilai ini kita boleh melukis garisan lurus yang terbaik bagi mewakili X dan Y. Persamaan regrasi yang akan dihasilkan adalah dalam fomat berikut:

$$y^{ram} = a + bx$$

Di mana y^{ram} merupakan nilai Y yang dikira menggunakan persamaan regrasi berdasarkan nilai X yang diberikan.

Dari persamaan (4), kita boleh bentukan dua persamaan yang baru iaitu:

$$\Sigma y_i = na + b\Sigma x_i \qquad \text{n adalah bilangan data yang dikumpul} \qquad (5)$$

$$\Sigma x_i y_i = a\Sigma x_i + b\Sigma x_i^2 \qquad (6)$$

Nilai a dan b merupakan anggaran bagi α dan ß. Nilai-nilai ini boleh didapati dengan menyelesaikan persamaan-persamaan berikut:

$$b = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2}$$

$$a = \frac{\Sigma y_i}{n} - b\left(\frac{\Sigma x_i}{n}\right) = \bar{y} - b\bar{x}$$

Dengan ini nilai a dan b sebenar adalah:

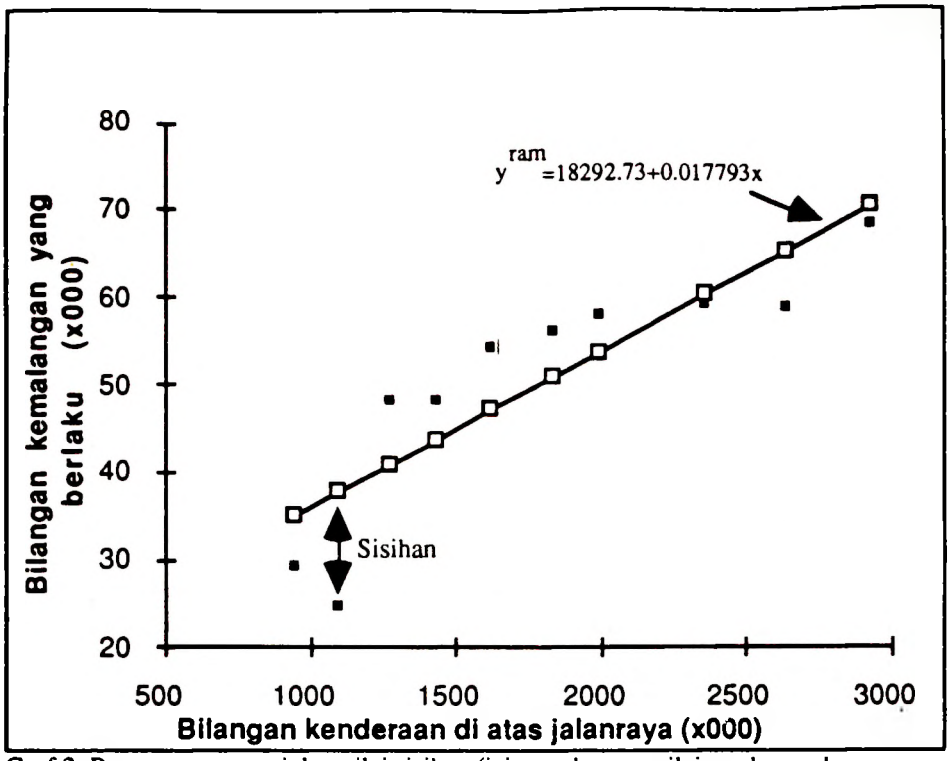
$$b = \frac{10(984337107826) - (18087249)(504747)}{10(36727116850415) - 18087249^2} = \frac{713886407257}{40122592116149} = \mathbf{0.017792629}$$

$$a = 50474.7 - (0.017792629)(1808724.9) = \mathbf{18292.72839}$$

Oleh itu persamaan regrasi yang menghubungkan pembolehubah bilangan kenderaan di atas jalan raya dengan pembolehubah bilangan kemalangan yang berlaku adalah seperti berikut:

$$\mathbf{y^{ram} = 18292.72839 + 0.017792629x} \qquad (7)$$

Dengan mengambil sebarang dua titik X yang sesuai dan gunakan persamaan (7) untuk mendapatkan nilai Y, kita boleh melukis persamaan regrasi tersebut di atas graf. Garislurus yang didapati adalah garislurus yang terbaik mewakili hubungan di antara X dan Y jika dibanding dengan garislurus-garislurus yang dibentuk menggunakan nilai a dan b yang lain. Tetapi, adakah garislurus ini mewakili keseluruhan data dengan baik? Perhatikan graf di bawah:



Graf 2: Persamaan regrasi dan nilai sisihan (iaitu perbezaan nilai y sebenar dengan nilai ramalan)

Kita tidak mungkin dapat melukis suatu garislurus yang melalui kesemua titik Y. Graf 2 menunjukkan tiada titik yang benar-benar berada di atas garisan regrasi. Walaubagaimana pun oleh kerana garislurus ini merupakan garislurus yang terbaik mewakili data, jumlah kuasadua sisihan (iaitu jarak tegak daripada setiap titik y_i kepada garisan kuasadua terkecil) bagi garisan ini adalah terkecil jika dibandingkan dengan sebarang garislurus-garislurus yang lain.

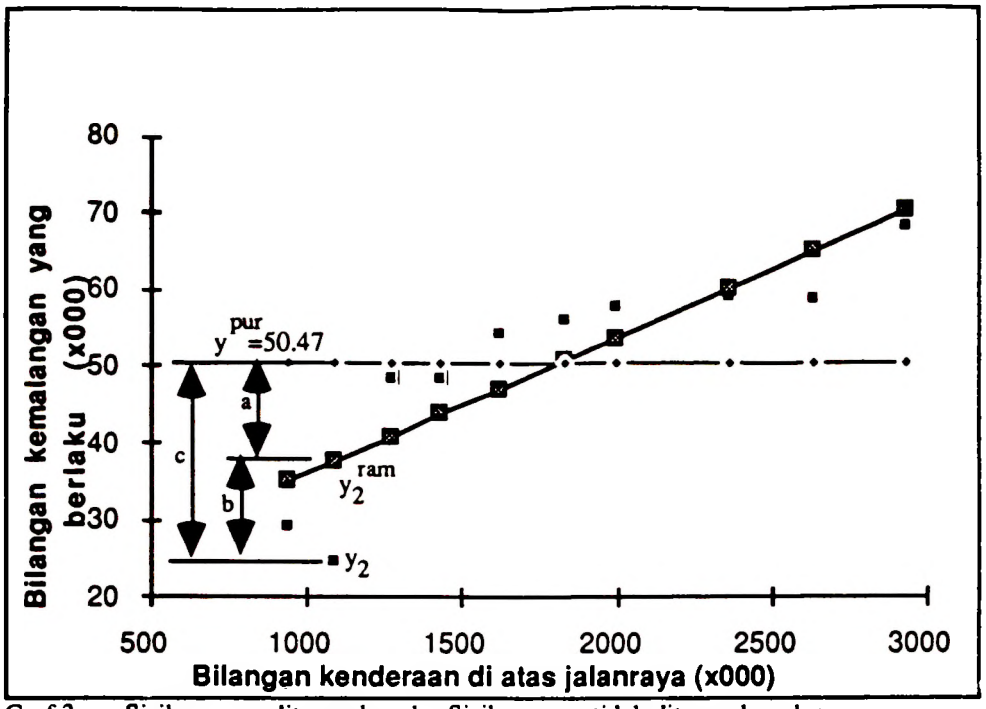
Salah satu ciri garisan kuasadua terkecil ini adalah jika kita set x_i bersamaan dengan x^{pur} (iaitu nilai purata bagi X), kita dapati nilai y_i juga akan bersamaan dengan y^{pur} (iaitu nilai purata bagi Y). Oleh itu garisan regresi ini melalui titik (x^{pur}, y^{pur}).

2.2 PENILAIAN PERSAMAAN REGRASI

Selepas mendapatkan persamaan regrasi, penilaian perlu dilakukan bagi menentukan kebagusan persamaan regrasi tersebut. Tujuan penentuan ini adalah untuk menentukan paras keyakinan terhadap nilai yang didapati dari proses ramalan dan anggaran yang akan dibincangkan nanti.

2.2.1 Jumlah Keseluruhan Kuasadua (Total Sum Squares)

Cara yang biasa digunakan untuk menilai kebagusan persamaan regrasi adalah dengan membanding serakan titik-titik Y berdasarkan garisan regrasi dan garisan y^{pur} (iaitu garisan purata y_i). Garisan y^{pur} adalah suatu garisan lurus yang mendatar kerana untuk sebarang nilai x, nilai y tetap konstan. Graf di bawah menunjukkan kedudukan garisan-garisan tersebut.



Graf 3: a - Sisihan yang diterangkan, b - Sisihan yang tidak diterangkan dan c - Jumlah sisihan untuk y_2

Perbezaan jarak dari titik y_i kepada garisan y^{pur} (iaitu $y_i - y^{pur}$) dipanggil ***jumlah sisihan*** (total deviation). Sebagai contohnya bagi y_2, jumlah sisihan adalah bersamaan dengan 24,581 - 50,474.7 =-25,893.7. Pengiraan untuk y_i yang lain juga adalah sama.

Jarak tegak daripada garisan regrasi kepada garisan y^{pur} bagi setiap titik Y (iaitu $y^{ram} - y^{pur}$) dipanggil sebagai ***sisihan yang diterangkan*** (explained deviation). Sebagai contohnya untuk y_2, sisihan yang diterangkan adalah 37,691.7 - 50,474.7 = -12,783.0.

Akhir sekali, jarak tegak di antara nilai y_i kepada garisan regrasi (iaitu $y_i - y^{ram}$) dipanggil ***sisihan yang tidak diterangkan*** (unexplained deviation). Contohnya bagi y_2 sisihan tidak diterangkan adalah 24,581.0 - 37,691.7 = -13,110.7. Hubungan di antara ketiga-tiga sisihan ini boleh ditakrifkan seperti berikut:

Jumlah sisihan = Sisihan yang diterangkan + sisihan yang tidak diterangkan

$$(y_i - y^{pur}) = (y^{ram} - y^{pur}) + (y_i - y^{ram}) \qquad (8)$$

Oleh itu bagi y_2 kita dapati -25,893.7 = -12,783.0 -13110.7.

Jika kita kuasaduakan setiap sisihan dan kita jumlahkan bagi kesemua nilai Y, hubungan bagi ketiga-tiga sisihan boleh ditulis seperti berikut:

$$\Sigma(y_i - y^{pur})^2 = \Sigma(y^{ram} - y^{pur})^2 + \Sigma(y_i - y^{ram})^2 \qquad (9)$$

Jumlah keseluruhan kuasadua = Jumlah kuasadua yang diterangkan + Jumlah kuasadua yang tidak diterangkan

Secara umumnya ketiga-tiga pengukuran ini mengukur serakan nilai-nilai Y. Jumlah keseluruhan kuasadua mengukur serakan nilai-nilai Y daripada nilai puratanya, y^{pur} atau dengan kata lain ia mengira jumlah variasi nilai-nilai Y (iaitu nilai jumlah kepembolehubahan data). Jumlah Kuasadua yang diterangkan pula mengira jumlah kepembolehubahan nilai-nilai Y yang diambilkira untuk hubungan linear antara nilai-nilai X dan Y (iaitu mengukur kepembolehubahan garisan regrasi). Jumlah kuasadua yang tidak diterangkan mengukur serakan nilai-nilai Y dari garisan regrasi (iaitu mengukur kepembolehubahan yang tidak diambil kira semasa garisan regrasi ditentukan). Nilai yang terakhir ini merupakan kuantiti yang diminimumkan sewaktu kita mengira bagi mendapatkan garisan kuasadua terkecil. Nilai ini juga dipanggil ***jumlah kuasadua ralat*** (error sum of squares).

2.2 Jadual Analisis Varian

Daripada andaian-andaian yang telah diberikan dahulu kita boleh menggunakan analisis varian untuk menguji kebagusan persamaan regrasi. Nilai darjah kebebasan (df) n-1 boleh dibahagikan kepada dua bahagian iaitu 1 untuk regrasi dan (n-1)-1 = n-2 untuk jumlah kuasadua kesalahan. Jika kita bahagikan jumlah kuasadua (SS) dengan nilai darjah kebebasan akan memberikan nilai ***purata kuasadua*** (mean squares, MS). Nilai F dengan darjah kebebasan n-2 dikira berdasarkan purata kuasadua regrasi (regression mean squares, MSR) dibahagikan dengan purata kuasadua kesalahan (error mean square, MSE).

Kita boleh menguji kewujudan regrasi linear pembolehubah bilangan kemalangan ke atas bilangan kenderaan di atas jalan raya mengunakan analisis varian seperti berikut:

Hipotesis yang hendak diuji adalah H_0: Tiada regrasi linear antara X dan Y (iaitu $\beta=0$) ataupun H_1: Terdapat regrasi linear Y ke atas X (iaitu $\beta\neq0$). Ujian statistik yang akan digunakan adalah F = MSR/MSE. Kita akan menggunakan taburan F dengan 1 dan 8 darjah kebebasan. Paras bererti (significance level) yang diperlukan adalah $\alpha = 0.05$. Peraturan membuat keputusan adalah: tolak H_0 jika nilai F yang dikira $\geq$ 5.32 (nilai dari jadual F). Pengiraan nilai F dilakukan seperti dalam jadual di bawah:

Sumber variasi	Jumlah kuasadua (SS)	Darjah kebebasan (df)	Purata kuasadua (MS)	F
Regrasi linear	1270191619 (= SSR)	1	1270191619 (= SSR/1)	24.13 (= MSR/MSE)
Sisihan dari kelinearan (kesalahan)	421157573 (= SSE)	8 (= n-2)	52644697 (= SSR/(n-2)	
Total	1691349192 (= SST)	9 (= n-1)		

Jadual 2: Jadual analisis varian untuk kajian kes

Oleh kerana 24.13 > 5.32, kita tolak H_0. Dengan ini kita boleh membuat kesimpulan bahawa berdasarkan dari sampel yang digunakan terdapat regrasi linear di antara kedua-dua pembolehubah. Oleh kerana 24.13 > 14.69 (sila rujuk jadual taburan F) maka $p < 0.005$.

2.3 Nilai R^2 dan R^2 Yang Diselaraskan

Nilai R^2 mengukur kebagusan garisan regrasi mewakili data. Kebagusan model tersebut perlu diketahui untuk mendapatkan paras keyakinan ke atas sebarang nilai yang diperolehi dari proses

ramalan dan anggaran. Secara mudahnya kita boleh menakrifkan R^2 sebagai nisbah jumlah kuasadua yang diterangkan kepada jumlah keseluruhan kuasadua.

$$R^2 = \frac{SSR}{SST} = \frac{\Sigma(y^{ram} - y^{pur})^2}{\Sigma(y_i - y^{pur})^2} \qquad (10)$$

Nilai ini menyatakan kadar jumlah variasi di dalam Y yang diterangkan oleh regrasi Y ke atas X. Nilai R^2 menjadi lebih bermakna jika dikira menggunakan formula berikut[1]:

$$R^2 = \frac{b^2 \Sigma (x_i - x^{pur})^2}{\Sigma (y_i - y^{pur})^2} = \frac{b^2 [\Sigma x_i^2 - (\Sigma x_i)^2 / n]}{\Sigma y_i^2 - (\Sigma y_i)^2 / n]} \qquad (11)$$

Nilai R^2 memberikan pengukuran tentang kehampiran persamaan regrasi kepada data yang digunakan. Lebih baik garisan regrasi mewakili data, nilai ini akan menghampiri 1. Dengan kata lain jika garisan regrasi melalui kesemua data dengan tepat, nilai R^2 akan bersamaan dengan 1. Ini adalah kerana y^{ram} dan y_i adalah sama. Dari persamaan (10) dan (11) kita dapat nilai pembahagi dan nilai kena bahagi adalah sama. Oleh itu nilai $R^2 = 1$. Nilai R^2 juga menjadi pengukuran tentang kelinearan data. Apabila persamaan regrasi mewakili data dengan baik, kedudukan data-data di atas graf adalah dalam garisan yang lurus dan apabila ini berlaku, nilai R^2 menghampiri kepada 1.

Apabila nilai darjah kebebasan kecil, R^2 mempunyai kecenderongan positif iaitu nilainya bertambah besar apabila lebih banyak pembolehubah ditambah ke dalam model. Oleh itu kita memerlukan pengukuran lain yang bebas dari kecenderongan ini. Pengukuran yang tidak berkencenderongan diberikan oleh nilai R^2 yang telah diselaraskan, R^2(sel). Formula yang digunakan untuk mengira nilai baru ini adalah:

$$R^2(sel) = 1 - \frac{\Sigma(y_i - y^{ram})^2 / (n-2)}{\Sigma(y_i - y^{pur})^2 / (n-1)}$$

Perbezaan antara R^2 dengan R^2(sel) adalah faktor (n-1)/(n-2). Apabila nilai n besar, faktor ini akan menghampiri 1 dan perbezaan antara kedua-dua nilai R^2 akan menghampiri 0. Nilai R^2 dan R^2(sel) bagi kes ini adalah:

$$R^2 = \frac{(0.017792629)^2 [36727116850415 - (18087249)^2 / 10]}{27168302593 - (504747)^2 / 10} = 0.75$$

$$R^2(sel) = 1 - \frac{421157573/8}{1691349192/9} = 0.72$$

Dengan ini regrasi Y ke atas X menerangkan 72% jumlah kepembolehubahan di dalam Y.

3.0 MENGGUNAKAN PERSAMAAN REGRASI

Apabila telah disahkan wujud hubungan linear antara X dan Y, kita boleh menggunakan persamaan regrasi bagi melakukan proses ***ramalan*** (prediction) dan ***anggaran*** (estimation).

Proses ramalan adalah dimana persamaan regrasi bagi nilai X tunggal digunakan untuk mendapat nilai Y. Proses anggaran pula menggunakan persamaan regrasi untuk mendapatkan nilai purata Y bagi nilai-nilai X yang sama. Dari segi pengiraan kedua-dua proses memberikan nilai Y yang sama tetapi dari segi nilai bagi selang ianya berbeza. Ini adalah kerana nilai anggaran purata pembolehubah tak bersandar kurang variasinya di bandingkan dengan nilainya X yang tunggal.

3.1 Meramal dan Menganggar Nilai Y untuk Nilai X Yang Di Ketahui

Kita boleh mendapatkan nilai ramalan Y bagi setiap nilai X, dengan menggantikan nilai tersebut kedalam persamaan regrasi. Selang keyakinan untuk proses ramalan ditakrifkan sebagai:

Ramalan ± (faktor kebolehpercayaan) x (ralat piawai ramalan).

Jika $\sigma^2_{y/x}$ tidak diketahui, selang ramalan 100(1- α)% untuk Y ditakrifkan sebagai:

$$y^{ram} \pm t_{1-\alpha/2}\, S_{y^{ram}}$$

di mana

$$S_{y^{ram}} = s_{y|x} \sqrt{1 + (1/n) + ((x_p - x^{pur})^2 / (\Sigma(x_i - x^{pur})^2))}$$

nilai $\Sigma(x_i - x^{pur})^2$ pula boleh diganti dengan $\Sigma x_i^2 - (\Sigma(x_i)^2) / n$, dengan darjah kebebasan n - 2 dan taburan t digunakan.

Sebagai contohnya katakan kita ingin meramalkan bilangan kemalangan yang akan berlaku jika terdapat 2,000,000 buah kenderaan di atas jalan raya. Nilai ramalan diperolehi dengan mengganti nilai x_p tersebut kedalam persamaan regrasi

$$y^{ram} = 18292.72839 + 0.017792629(2000000) = 53878.$$

Oleh itu sebanyak 53878 bilangan kemalangan diramalkan akan berlaku jika terdapat 2,000,000 kenderaan di atas jalan raya.

95 % selang ramalan diberikan oleh:

$$53878 \pm 2.306\ X$$

$$(\sqrt{52644697})\ \sqrt{(1 + (1/10) + ((2000000 - 50474.7)^2) / (36727116850415 - (18087249)^2 / 10)}$$

$$53878 \pm 17620.81 \quad = \quad (36257,\ 71498)$$

Oleh itu kita mempunyai keyakinan sebanyak 95% nilai ramalan kemalangan sebenar bagi 2,000,000 bilangan kenderaan di atas jalan raya akan berada dalam selang ini.

Untuk mengganggarkan purata $\mu_{y|x}$ bagi subpopulasi Y untuk nilai X yang tertentu, prosesnya adalah sama iaitu dengan menggantikan x_p kedalam persamaan regrasi.

Selang keyakinan untuk $\mu_{y|x}$ ditakrifkan seperti berikut:
Selang keyakinan 100(1-α)% untuk $\mu_{y|x}$, apabila $\sigma^2_{y|x}$ tidak diketahui diberikan oleh

$$\mu^{ram}_{y|x} \pm t_{1-\alpha/2} S_{\mu^{ram}_{y|x}}$$

di mana

$$S_{\mu^{ram}_{y|x}} = s_{y|x} \sqrt{(1/n) + ((x_p - x^{pur})^2 / (\Sigma(x_i - x^{pur})^2))}$$

Katakan kita ingin menganggar purata bilangan kemalangan bagi purata bilangan kenderaan 2,000,000 buah.

$$\mu^{ram}_{y|x} = 18292.72839 + 0.017792629(2000000) = 53878$$

Oleh itu purata kemalangan sebanyak 53878 kali akan berlaku sekira terdapat purata 2,000,000 buah kenderaan di atas jalan raya.

95% selang keyakinan untuk $\mu^{ram}_{y|x}$ adalah:

$$53878 \pm 2.306 \text{ X}$$

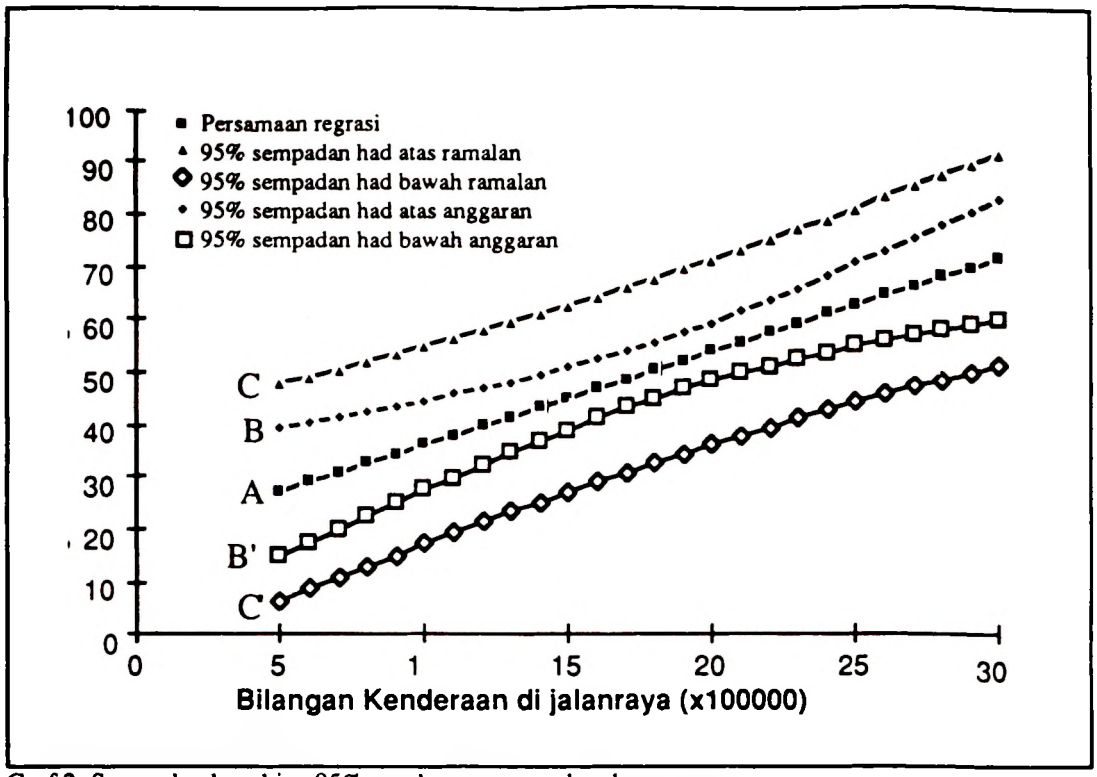
$$(\sqrt{52644697}) \quad \sqrt{((1/10) + ((2000000 - 50474.7)^2) / (36727116850415 - (18087249)^2 / 10)}$$

$$53878 \pm 5526.96 = (48351, 59404)$$

Oleh itu jika beberapa sampel yang terdiri dari 2,000,000 kenderaan di atas jalan raya diambil, kemalangan yang akan berlaku dianggarkan sebanyak 53878 dan dengan 95% selang keyakinan bilangan kemalangan yang dijangka akan berlaku akan berada di antara 48351 hingga 59404.

3.2 Membina Sempadan Keyakinan

Katakan kita membina sempadan keyakinan untuk nilai-nilai ramalan dan anggaran berdasarkan 95% selang keyakinan. Graf bagi sempadan keyakinan tersebut ditunjukan di bawah.



Graf 2: Sempadan keyakian 95% untuk proses ramalan dan anggaran

Garislurus di bahagian tengah (iaitu garisan A) merupakan garisan regrasi. Lengkuk yang terdekat dengan garisan regrasi di sebelah atas dan bawah (iaitu garisan B dan B') adalah lengkuk 95% selang keyakinan proses anggaran. Oleh itu 95% selang keyakian anggaran akan berada di antara kedua-dua lengkuk ini. Lengkuk ini dilukis dengan menyambung titik yang dikira bagi setiap x_p yang diberikan. Lengkuk C dan C' pula merupakan 95% selang keyakinan untuk proses ramalan. Di sini dapat diperhatikan bahawa selang keyakinan bagi proses ramalan lebih besar dari proses anggaran dan kedua-dua selang ini terkecil apabila $x_p = x^{pur.}$ Apabila nila x_p bertambah atau berkurang selang akan bertambah besar.

4.0 ANALISIS REGRASI MUDAH MENGGUNAKAN SAS

Untuk memudahkan proses mendapat persamaan regrasi, kita boleh menggunakan pakej komputer seperti SAS/STAT menggunakan prosidur REG. Aturcara SAS/STAT untuk melakukan analisis di atas ditunjukkan dalam Lampiran 1.

Data-data yang dibaca disimpan dalam set data SAS yang dinamakan "Kemal". Tiga pembolehubah yang terlibat iaitu tahun, bilangan kenderaan (b_kend) dan bilangan kemalangan (b_kemal) yang berlaku. Selepas pernyataan CARDS, data-data kemalangan dari tahun 1973 hingga 1982 dimasukkan dan diakhiri dengan tanda ";". Prosidur REG dipanggil untuk melakukan proses analisis regrasi ke atas data yang berada dalam set data SAS Kemal. Pernyataan MODEL digunakan untuk menakrifkan model yang ingin dibentuk. Dalam pernyataan ini juga pembolehubah bersandar dan pembolehubah-pembolehubah tak bersandar ditakrifkan. Pembolehubah bersandar mesti diletakkan sebelum tanda "=" manakala pembolehubah tak bersandar pula diletakkan selepasnya. Pilihan P dan R digunakan supaya nilai Reja dan nilai Ramalan (Predict) dikeluarkan di dalam laporan nanti. Nilai ramalan adalah nilai-nilai yang berada di atas garisan regrasi bagi setiap bilangan kenderaan di atas jalan raya. Nilai reja pula adalah nilai perbezaan di antara nilai Y yang sebenar dengan nilai ramalan. Pernyataan TITLE digunakan untuk menakrifkan tajuk yang akan dicetak disetiap muka laporan. Arahan RUN digunakan untuk memulakan proses analisis seperti yang telah ditakrifkan. Perhatikan setiap arahan SAS mesti berakhir dengan tanda semikolon.

Output SAS/STAT ditunjukkan dalam Lampiran 2. Nilai bagi a dan b boleh didapati di bawah lajur Parameter Estimate. Nilai INTERCEPT adalah nilai bagi a manakala nila bagi b diberikan oleh B_KEND.

Nilai Prob>|T| menceritakan tentang kebagusan model yang telah dibentuk. Oleh kerana nilainya kecil (iaitu 0.0012), jadi kita boleh membuat kesimpulan bahawa model itu keseluruhannya baik iaitu terdapat regrasi linear Y ke atas X. Nilai R^2 dan R^2(sel) yang didapati dari SAS/STAT juga menyokong kesimpulan di atas.

5.0 KESIMPULAN

Model Regrasi Mudah boleh digunakan jika hanya terdapat satu pembolehubah tak bersandar sahaja. Walaubagaimana pun di dalam keadaan sebenar terdapat beberapa pembolehubah tak bersandar di dalam model yang hendak dibina. Model yang terdiri dari beberapa pembolehubah tak bersandar dipanggil model regrasi berbagai (multiple regression model). Pengiraan bagi model kedua ini lebih kompleks tetapi dengan bantuan komputer prosesnya menjadi mudah dan cepat. Kertas kerja bagi model regrasi berbagai akan dikeluarkan tidak lama lagi.

Di dalam analisis yang telah dilakukan didapati terdapat hubungan langsung di antara bilangan kenderaan di atas jalan raya dengan bilangan kemalangan yang berlaku. Dengan ini untuk mengurangkan bilangan kemalangan yang berlaku, pihak tertentu sepatutnya berusaha untuk mengurangkan bilangan kenderaan di atas jalan raya. Ini dapat dilakukan mungkin dengan meningkatkan taraf pengangkutan awam, menaikan kadar bayaran tol serta bayaran letak kereta dan sebagainya. Dengan usaha-usaha ini diharap orang ramai akan menggunakan pengangkutan awam yang telah disediakan dan dijangka kadar kemalangan akan berkurangan.

RUJUKAN

1. Wayne W. Daniel, James C. Terrell, Business Statistics For Management and Economics, fifth edition, Houghton Mifflin Company, Boston

2. Statistical Report Road Accidents Malaysia, Royal Malaysia Police, 1982

3. SAS/STAT User Guide, Version 6.

4. SAS/BASIC User Guide, Version 6.

5. J. Supranto (1986), Kaedah Penyelidikan Penggunaannya Dalam Pemasaran, Dewan Bahasa dan Pustaka.

6. Richard I. Levin and David S. Rubin, Statistics For Management, Sixth Edition, Prentice Hall International Editions.

LAMPIRAN 1

```
DATA kemal;
INPUT tahun b_kend b_kemal;
CARDS;
1973    939951    29286
1974    1090279   24581
1975    1267119   48233
1976    1429845   48291
1977    1621271   54222
1978    1829958   56021
1979    1989391   57931
1980    2357386   59084
1981    2631948   58768
1982    2930101   68330
;
PROC REG DATA=kemal;
MODEL b_kemal=b_kend / P R;
TITLE 'ANALISIS KEMALANGAN JALAN RAYA';
RUN;
```

LAMPIRAN 2

ANALISIS KEMALANGAN JALAN RAYA

12:35 Saturday, February 6, 1993

Model: MODEL1
Dependent Variable: B_KEMAL

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	1270191619.2	1270191619.2	24.128	0.0012
Error	8	421157572.90	52644696.612		
C Total	9	1691349192.1			

Root MSE	7255.66652	R-square	0.7510	
Dep Mean	50474.70000	Adj R-sq	0.7199	C.V.14.37486

ANALISIS KEMALANGAN JALAN RAYA

12:35 Saturday, February 6, 1993

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	18293	6941.8662304	2.635	0.0299
B_KEND	1	0.017793	0.00362229	4.912	0.0012

ANALISIS KEMALANGAN JALAN RAYA

12:35 Saturday, February 6, 1993

Obs	Dep Var B_KEMAL	Predict Value	Std Err Predict	Residual	Std Err Residual	Student Residual
1	29286.0	35016.9	3894.580	-5730.9	6121.842	-0.936
2	24581.0	37691.7	3469.444	-13110.7	6372.413	-2.057
3	48233.0	40838.1	3018.830	7394.9	6597.830	1.121
4	48291.0	43733.4	2673.571	4557.6	6745.125	0.676
5	54222.0	47139.4	2392.807	7082.6	6849.757	1.034
6	56021.0	50852.5	2295.732	5168.5	6882.900	0.751
7	57931.0	53689.2	2385.947	4241.8	6852.150	0.619
8	59084.0	60236.8	3035.500	-1152.8	6590.177	-0.175
9	58768.0	65122.0	3762.512	-6354.0	6203.886	-1.024
10	68330.0	70426.9	4665.177	-2096.9	5557.052	-0.377

Obs	-2-1-0 1 2	Cook's D
1	*| |	0.177
2	****| |	0.627
3	|** |	0.131
4	|* |	0.036
5	|** |	0.065
6	|* |	0.031
7	|* |	0.023
8	| |	0.003
9	**| |	0.193
10	| |	0.050

Sum of Residuals 0
Sum of Squared Residuals 421157572.90
Predicted Resid SS (Press) 670521508.77