Penerapan Imputasi Ganda dengan Metode *Predictive Mean Matching* (PMM) untuk Pendugaan Data Hilang Pada Model Regresi Linear

Wilsen[1, a)], Widyanti Rahayu[2, b)], Vera Maya Santi[2, c)]

[1]*Program Studi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta*
[2]*Program Studi Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Jakarta*

Email: [a)]wilsen1996@gmail.com, [b)]widyanti.rahayu@gmail.com, [c)]vmsanti@unj.ac.id

Abstract

Missing data has been a common problem in linear regression analysis. However, many researchers ignore that problem. As a result, the power to find a significant result decreases. In addition, the lost data can cause bias in the estimation of parameters. Some statistical analyzes were developed to cope with missing data in linear regression model. One of multiple imputation method is Predictive Mean Matching (PMM) as used in this research. Studies in this research focused to overcome missing covariate data that follows the pattern of monotone and mechanisms of Missing At Random (MAR) on the data. Automobile dataset is used to see how the effectiveness of the use PMM is analyzed based on the relative efficiency estimation result of missing covariate data. Compression ratio, bore, and stroke are used as the independent variables and fuel efficieny (city mpg) as the dependent variable. Missing data occurs in bore and stroke. The percentage of missing data is 2%. The result showed that PMM just loss of 5% relative efficieny. Moreover, compression ratio, bore, and stroke are significant based on the Wald test and t-test at 5% level of significance (α=5%).

Keywords: missing data, linear regression, multiple imputation, predictive mean matching, relative efficieny, wald test, t-test.

Abstrak

Permasalahan data hilang pada analisis regresi linear telah menjadi hal yang umum. Meskipun demikian, beberapa peneliti mengabaikan masalah data hilang. Akibatnya, tingkat ketepatan model yang diperoleh akan menurun. Bahkan, dugaan parameter yang dihasilkan cenderung bias. Salah satu teknik yang digunakan untuk mengatasi masalah data hilang pada model regresi linear adalah imputasi ganda dengan metode Predictive Mean Matching (PMM). Kajian dalam tulisan ini difokuskan untuk mengatasi permasalahan data hilang yang terdapat pada variabel bebas dengan pola monoton dan mekanisme Missing At Random (MAR). Imputasi ganda dengan metode PMM diterapkan pada kasus tingkat efisiensi bahan bakar (City mpg). Variabel bebas yang digunakan dalam kasus tersebut adalah Compression ratio, Bore, dan Stroke. Persentase data hilang sebesar 2% yang tersebar di variabel Bore dan Stroke. Hasil data imputasi memenuhi batas toleransi dari nilai efisiensi relatif sebesar 5%. Bahkan, Compression ratio, Bore, dan Stroke memberikan pengaruh yang signifikan terhadap tingkat efisiensi bahan bakar (City mpg) berdasarkan Uji Wald dan Uji t dengan taraf signfikansi sebesar 5% (α=5%).

Kata-kata kunci: data hilang, regresi linear, imputasi ganda, predictive mean matching, efisiensi relatif uji wald, uji t.

PENDAHULUAN

Analisis regresi digunakan untuk mengetahui hubungan di antara beberapa variabel (Yan & Su 2009, h.1). Ada dua variabel dalam analisis regresi, yaitu variabel terikat (Y) dan variabel bebas (X). Hubungan antara variabel terikat dan variabel bebas dalam analisis regresi digambarkan melalui model matematis yang dikenal sebagai persamaan regresi. Persamaan regresi yang menunjukkan hubungan linear antara satu variabel terikat (Y) dan satu variabel bebas (X_1) dikenal sebagai regresi linear sederhana, sedangkan persamaan regresi yang memodelkan hubungan linear antara satu variabel terikat dengan lebih dari satu variabel bebas disebut sebagai regresi linear berganda.

Berbagai permasalahan seringkali ditemukan dalam model regresi linear. Masalah tersebut pada umumnya berkaitan dengan pelanggaran asumsi klasik, seperti normalitas, multikolinearitas, autokorelasi atau heteroskedastisitas. Akan tetapi, permasalahan yang ditemukan dalam model regresi linear juga dapat terjadi pada proses pengumpulan data di lapangan. Salah satu contoh masalah tersebut adalah adanya data yang hilang (*missing data*). Akibatnya, jumlah sampel yang ideal gagal didapatkan. Kondisi tersebut diperparah oleh keterbatasan biaya dan waktu, sehingga kegiatan pengumpulan data tidak dapat diulang oleh peneliti. Ketika metode kuadrat terkecil digunakan untuk menduga parameter tersebut, penduga yang didapatkan cenderung bias (Moris dkk, 2014). Hal tersebut dikarenakan jumlah sampel yang diperoleh tidak cukup untuk mewakili populasi. Oleh karena itu, data yang hilang pada model regresi linear membutuhkan penanganan khusus.

Salah satu teknik yang digunakan untuk mengatasi data hilang pada model regresi linear adalah metode imputasi tunggal yang dikenalkan oleh Rubin di tahun 1970-an. Metode tersebut dilakukan dengan cara mengisi nilai yang hilang dengan satu nilai tertentu. Selanjutnya, Rubin di tahun 1988 mengembangkan metode imputasi tunggal menjadi imputasi ganda. Metode imputasi ganda menghasilkan beberapa nilai dugaan yang mewakili distribusi kemungkinannya dan dilakukan sebanyak *m* kali (Buuren 2012, h. 26). Metode tersebut lebih efektif dibandingkan dengan imputasi tunggal, karena nilai dugaan yang dihasilkan lebih bervariasi. Berdasarkan uraian tersebut, teknik imputasi ganda dengan Metode *Predictive Mean Matching* (PMM) digunakan dalam pendugaan data hilang pada model regresi linear.

METODE

Regresi Linear Berganda

Model regresi linear berganda menurut Yan dan Su (2009, h.41) menunjukkan hubungan linear antara satu variabel terikat (Y) dengan lebih dari satu variabel bebas ($X_1, X_2, \cdots, X_n$), yaitu:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i \; ; i = 1, 2, \ldots, n \qquad (1)$$

Variabel terikat Y berhubungan dengan variabel bebas p, di mana Y_i adalah variabel terikat pada pengamatan ke-i, $X_{i1}, X_{i2}, \cdots, X_{in}$ adalah nilai variabel bebas pada pengamatan ke-*i* yang terkait dengan parameter ke-p, sedangkan $\beta_0, \beta_1, \cdots, \beta_k$ adalah parameter atau koefisien regresi dan ε_i adalah error pengamatan ke-*i*. Parameter $\beta_0, \beta_1, \cdots, \beta_p$ pada model regresi linear berganda $Y = X\beta + \varepsilon$ akan diduga menggunakan Metode Kuadrat Terkecil (*Ordinary Least Square*), yaitu:

$$\hat{\beta} = \left(X'X \right)^{-1} X'Y \qquad (2)$$

Data Hilang

Metode statistika pada umumnya dibentuk untuk menganalisis data lengkap. Akan tetapi, data yang lengkap terkadang tidak selalu tersedia dalam kehidupan sehari-hari. Kejadian tersebut biasanya dikenal dengan istilah data hilang. Menurut Little dan Rubin (2002, h.3) salah satu penyebab data

hilang adalah adanya *non-response*, seperti responden tidak mengetahui atau menolak untuk menjawab. Akibatnya, tujuan penelitian menjadi tidak representatif dan berbias.

Pola Data Hilang

Menurut Buuren (2012, h. 95), pola data hilang terdiri dari:

1. Univariat dan Multivariat. Data hilang dikatakan memiliki pola univariat apabila hanya ada satu variabel yang mengalami masalah data hilang
2. Monoton dan *non-monoton* (umum). Data hilang dikatakan berpola monoton ketika data yang hilang pada pengukuran tertentu selalu hilang pada pengukuran berikutnya. Ketika pola monoton tidak terpenuhi, data hilang disebut *non-monoton* (umum).
3. Terhubung dan tidak terhubung. Pola terhubung terjadi apabila data hasil observasi dapat diakses dari observasi yang lainnya dengan cara berpindah secara vertikal atau horizontal. Ketika antardata hasil observasi tidak dapat dihubungkan, baik dengan perpindahan vertikal atau horizontal, data hilang dikatakan memiliki pola tidak terhubung.

Mekanisme Data Hilang

Buuren (2012, h. 6) menunjukkan bahwa ada tiga mekanisme data hilang, yaitu:

1. MCAR (*Missing Completely At Random*)
   Mekanisme data hilang secara MCAR tidak memiliki keterkaitan (saling bebas), baik dengan variabel yang diamati ataupun yang tidak diamati. Oleh karena itu, mekanisme data hilang secara MCAR terjadi apabila besarnya peluang suatu data akan hilang adalah sama dan acak.
2. MAR (*Missing At Random*)
   Data yang hilang dengan mekanisme MAR tidak selalu terjadi secara acak seperti pada mekanisme MCAR, melainkan bergantung pada data hasil observasi. Hal tersebut dapat dilihat berdasarkan nilai peluang suatu data menjadi hilang. Nilai peluang suatu data akan hilang dalam mekanisme MAR tidak memiliki bobot yang sama, melainkan bergantung dengan hasil pengukuran dari observasi lain yang diteliti.
3. MNAR (*Missing Not At Random*)
   Mekanisme MNAR berbeda dengan mekanisme MCAR ataupun MAR. Proses hilangnya suatu data tidak hanya bergantung dengan data yang telah diobservasi, tetapi bergantung juga dengan berbagai faktor di luar pengukuran yang dilakukan.

Metode Imputasi

Imputasi merupakan salah satu metode yang digunakan untuk mengatasi data hilang pada model regresi linear. Metode tersebut bekerja dengan cara mengisi nilai yang hilang dengan nilai tertentu. Nilai yang dihasilkan dari proses imputasi akan diperlakukan seperti data riil. Metode imputasi secara umum dibedakan menjadi dua, yaitu imputasi tunggal (*single imputation*) dan imputasi ganda (*multiple imputation*).

Imputasi Tunggal

Imputasi tunggal adalah kegiatan pengisian data hilang dengan satu nilai tertentu yang diperoleh dari perhitungan sederhana. Dengan kata lain, imputasi tunggal tidak memiliki proses repetisi. Salah satu contoh imputasi tunggal adalah imputasi rata-tata (*mean imputation*). Hasil imputasi diperoleh dari perhitungan rata-rata pada kumpulan data lengkap yang ada di variabel tersebut.

Imputasi Ganda

Imputasi ganda merupakan pengembangan dari metode imputasi tunggal. Teknik tersebut menghasilkan beberapa nilai dugaan yang mewakili distribusi kemungkinannya dan dilakukan

sebanyak *m* kali (Buuren 2012, h. 26). Tahapan dalam menggunakan teknik imputasi ganda dibedakan menjadi tiga, yaitu proses imputasi, analisis dan kombinasi.

Predictive Mean Matching (PMM)

Salah satu pendekatan yang dapat digunakan dalam pendugaan data hilang adalah pengambilan sampel acak dari data yang berhasil diobservasi berdasarkan selisih jarak yang minimum. Metode dalam imputasi ganda yang memanfaatkan pendekatan tersebut adalah metode *Predictive Mean Matching* (PMM). Adapun, tahapan metode PMM dalam imputasi ganda terdiri atas tahap imputasi, analisis dan kombinasi.

Tahap Imputasi

Proses imputasi awal yang digunakan sebanyak 5 kali. Oleh karena itu, ada lima kelompok data hasil imputasi yang diperoleh. Langkah-langkah dalam penggunaan metode PMM adalah sebagai berikut.

1. Mendapatkan nilai V melalui rumus $V = (X'X)^{-1}$
2. Menduga parameter regresi $\hat{\beta}$ menggunakan rumus kuadrat terkecil, yaitu $\hat{\beta} = (X'X)^{-1} X'Y_{obs}$
3. Menghitung nilai parameter $\bar{\sigma}^{2*}$ yang diperoleh melalui persamaan $\sigma^{2*} = \dfrac{(Y_{obs} - X\hat{\beta})'(Y_{obs} - X\hat{\beta})}{g}$

   g merupakan variabel acak yang berdistribusi *chi square* dengan derajat bebas *n-p*. Banyaknya hasil observasi adalah *n*, sedangkan *p* melambangkan banyaknya variabel bebas ditambah satu.
4. Menduga parameter $\hat{\beta}*$ dengan menggunakan persamaan $\hat{\beta}* = \hat{\beta} + \hat{\sigma} * L_{ij}' z$

   L_{ij}' adalah matriks segitiga atas yang dihasilkan dari dekomposisi cholesky $V = L_{ij} L_{ij}'$, sedangkan z adalah vektor berukuran $p \times 1$ yang anggota-anggotanya berdistribusi normal baku.
5. Menghitung selisih nilai prediksi untuk data yang hilang, yaitu $\eta\,(i, j) = \left| X_{obs,[i]}\hat{\beta} - X_{mis,[j]}\hat{\beta}* \right|$ dengan $i = 1, \cdots, n$ dan $j = 1, \cdots, n_0$. X_{obs} merupakan matriks prediktor berukuran $n \times p$ dengan data yang berhasil diobservasi pada Y, sedangkan X_{mis} merupakan matriks prediktor berukuran $n_0 \times p$ dengan data yang hilang pada Y.
6. Mengumpulkan selisih nilai prediksi data hilang sebanyak n_0 himpunan. Setiap himpunan n_0 berisikan *d* data.
7. Memilih *d* data lengkap yang memiliki selisih jarak minimum antarnilai prediksi. Pemilihan *d* berdasarkan hasil $\left| X_{obs,[i]}\hat{\beta} - X_{mis,[j]}\hat{\beta}* \right|$ yang minimum. Nilai *d* dalam tulisan ini adalah 5 (Buuren 2012).
8. Memilih salah satu data secara acak dari *d* data yang tersedia untuk menggantikan data yang hilang.

Tahap imputasi ganda dengan PMM menggunakan proses imputasi awal sebanyak 5 (*m=5*), sehingga tahap imputasi diulang sebanyak 5 kali.

Tahap Analisis

Analisis yang digunakan dalam tulisan ini adalah analisis regresi linear berganda. Akibatnya, setiap kumpulan data yang yang berhasil diperoleh dari proses imputasi masing-masing akan memiliki penduga parameter regresi. Dengan demikian, terdapat 5 himpunan penduga parameter regresi dalam proses tersebut.

Tahap Kombinasi

Tahap kombinasi merupakan tahapan terakhir dalam imputasi ganda. Tahap tersebut akan memperoleh himpunan penduga parameter regresi yang tunggal. Himpunan tersebut didapatkan dengan cara menghitung rata-rata dari *m* penduga parameter yang telah diperoleh dalam tahap analisis. Tahap kombinasi dibahas dengan rinci oleh White dkk (2010, h. 378).

Evaluasi Data Hasil Imputasi

Evaluasi data hasil imputasi dapat dilihat berdasarkan nilai efisiensi relatif dan uji kelayakan model.

Nilai Efisiensi Relatif

Efisiensi relatif (ER) adalah efisiensi yang diperoleh dengan menggunakan *m* imputasi terbatas dibandingkan dengan jumlah tidak terbatas. Data hasil imputasi akan semakin baik apabila nilai efisiensi relatif semakin mendekati satu. Formulasi untuk mendapatkan nilai efisiensi relatif adalah sebagai berikut.

$$ER = \left(1 + \frac{\gamma}{m}\right)^{-1} \tag{3}$$

Nilai efisiensi relatif bergantung dengan besarnya fraksi data hilang (γ) dan banyaknya proses imputasi (*m*). Nilai tersebut akan semakin baik apabilai fraksi data hilang mendekati nol, sedangkan banyaknya proses imputasi mendekati tak hingga. Penjelasan lebih lanjut terkait dengan efisiensi relatif dapat ditemukan di White dkk. (2010, h. 387).

Uji Kelayakan Model

Uji kelayakan model yang digunakan terdiri dari uji wald dan uji t. Tingkat signifikansi yang digunakan adalah 5% (α=5%).

1. Uji Wald
   Signifikansi semua variabel bebas terhadap variabel terikat secara bersama-sama (simultan) dapat diketahui berdasarkan Uji Wald. Oleh karena itu, uji tersebut dapat mengetahui kelayakan suatu model regresi yang datanya berasal dari proses imputasi ganda. Penjelasan terkait dengan uji wald dapat ditemukan di Marshall dkk. (2009, h.4).
2. Uji t
   Uji t menentukan signifikansi dari masing-masing variabel bebas terhadap variabel terikat.

Sumber Data

Sumber data yang digunakan adalah *Automobile Dataset*. Data tersebut diperoleh dari *Machine Learning Repository* (Schlimmer, 1987, h. 1). Pemilihan variabel terikat dan variabel bebas didasarkan oleh penelitian Ahmad dkk (2014, h. 1). Adapun, rincian variabel terikat dan variabel bebas terhadap data tersebut adalah sebagai berikut.

1. Variabel terikat: *City mpg* (*miles per gallon*)
2. Variabel bebas yang digunakan terdiri atas *Compression ratio* (%), *Bore* (in), dan *Stroke* (in)

Data tersebut mengalami data hilang pada variabel *Bore* dan *Stroke* sekitar 2% dengan pola data hilang monoton.

HASIL DAN PEMBAHASAN

Analisis Data

Pendugaan data hilang pada model regresi linear menggunakan imputasi ganda dengan metode PMM terdiri menjadi tiga tahapan, yaitu tahap imputasi, analisis, dan kombinasi.

Tahap Imputasi

Jumlah proses awal imputasi yang digunakan dalam tulisan ini adalah lima (m =5), sehingga menghasilkan lima kelompok data yang saling berbeda. Adapun, hasil data imputasi berdasarkan *software-R* dengan package *mice* ditampilkan di Tabel 1 s.d. 5.

TABEL 1. DATA HASIL IMPUTASI PERTAMA

Data Ke-	*Bore* (X_2)	*Stroke* (X_3)
56	3.47	2.64
57	3.43	3.07
58	3.62	3.50
59	3.74	3.15

TABEL 2. DATA HASIL IMPUTASI KEDUA

Data Ke-	*Bore* (X_2)	*Stroke* (X_3)
56	3.74	3.11
57	3.62	3.29
58	3.80	2.64
59	3.46	3.23

TABEL 3. DATA HASIL IMPUTASI KETIGA

Data Ke-	*Bore* (X_2)	*Stroke* (X_3)
56	3.58	2.80
57	3.46	3.12
58	3.58	3.23
59	3.74	3.23

TABEL 4. DATA HASIL IMPUTASI KEEMPAT

Data Ke-	*Bore* (X_2)	*Stroke* (X_3)
56	3.46	2.80
57	3.74	3.50
58	3.46	2.80
59	3.46	2.68

TABEL 5. DATA HASIL IMPUTASI KELIMA

Data Ke-	*Bore* (X_2)	*Stroke* (X_3)
56	3.94	2.64
57	3.94	2.64
58	3.43	3.27
59	3.62	3.29

Variasi antarimputasi disebabkan karena teknik imputasi ganda memperhitungkan faktor ketidakpastian (*uncertainty*) dari sampel. Hal tersebutlah yang menunjukkan kelebihan dari teknik imputasi ganda. Oleh karena itu, metode PMM dalam menangani kasus *Automobile Dataset* menghasilkan lima himpunan data. Himpunan tersebut akan digunakan dalam tahap analisis. Dengan kata lain, masing-masing kelompok data tersebut akan memiliki himpunan penduga parameter.

Tahap Analisis

Tahap analisis merupakan lanjutan dari tahap imputasi. Oleh karena itu, tahap analisis juga akan menghasilkan lima analisis yang saling berbeda. Hal tersebut dikarenakan terdapat lima kelompok data dari proses imputasi. Analisis yang digunakan pada kasus tersebut adalah analisis regresi linear

berganda. Hasil pendugaan parameter yang diperoleh dari masing-masing kelompok data imputasi disajikan pada Tabel 6 dengan *i* menyatakan hasil imputasi ($i = 1, 2, \cdots, 5$).

TABEL 6. HASIL ANALISIS DATA IMPUTASI

Penduga Parameter	Himpunan Data (*i*)				
	1	**2**	**3**	**4**	**5**
$\hat{\beta}_{0i}$	60.791	62.256	61.945	57.801	61.969
$\hat{\beta}_{1i}$	2.106	2.093	2.085	2.170	2.112
$\hat{\beta}_{2i}$	-13.948	-14.070	-14.051	-13.759	-14.002
$\hat{\beta}_{3i}$	-2.595	-2.877	-2.785	-2.047	-2.905

Tahap Kombinasi

Tahap kombinasi merupakan tahapan terakhir dalam imputasi ganda. Tahap tersebut akan memperoleh himpunan penduga parameter regresi yang tunggal dengan cara menghitung rata-rata dari *m* penduga parameter yang telah diperoleh dalam tahap analisis. Contoh kasus *Automobile Dataset* yang digunakan dalam pembahasan menghasilkan lima kelompok penduga parameter pada tahap analisis, sehingga nilai $\bar{Q}$ diperoleh berdasarkan rata-rata dari kelima kelompok penduga parameter $\hat{Q}_i$. Hasil keseluruhan penduga parameter $\hat{Q}_i$ ditampilkan pada Tabel 7.

TABEL 7. RATA-RATA PENDUGA PARAMETER HASIL DATA IMPUTASI

Penduga Parameter	Himpunan Data (*i*)					
	1	**2**	**3**	**4**	**5**	**Rata-rata**
	$\hat{Q}_1$	$\hat{Q}_2$	$\hat{Q}_3$	$\hat{Q}_4$	$\hat{Q}_5$	$\bar{Q}$
$\hat{\beta}_{0i}$	60.791	62.256	61.945	57.801	61.969	60.952
$\hat{\beta}_{1i}$	2.106	2.093	2.085	2.170	2.112	2.113
$\hat{\beta}_{2i}$	-13.948	-14.070	-14.051	-13.759	-14.002	-13.966
$\hat{\beta}_{3i}$	-2.595	-2.877	-2.785	-2.047	-2.905	-2.642

Model regresi yang diperoleh berdasarkan hasil tahap kombinasi adalah sebagai berikut.

$$Y_i = 60.952 + 2.113X_{i1} - 13.966X_{i2} - 2.642X_{i3} \quad (4)$$

Interpretasi Model

Interpretasi koefisien regresi yang terdapat di Persamaan (4) adalah sebagai berikut.

1. Koefisien X_1 sebesar 2.113 menyatakan bahwa peningkatan satu persen dari *Compression ratio*, akan meningkatkan nilai *City mpg* sebesar 2.113 satuan.
2. Koefisien X_2 sebesar -13.966 menyatakan bahwa peningkatan satu inci dari *Bore*, akan menurunkan nilai *City mpg* sebesar -13.966 satuan.
3. Koefisien X_3 sebesar -2.642 menyatakan bahwa peningkatan satu inci dari *Stroke*, akan menurunkan nilai *City* mpg sebesar -2.642 satuan.

Evaluasi Data Hasil Imputasi

Evaluasi data hasil imputasi menggunakan metode PMM dapat dilihat berdasarkan kriteria nilai efisiensi relatif dan uji kelayakan model regresi. Hasil efisiensi relatif dari masing-masing penduga parameter ditampilkan dalam Tabel 8. Sementara, hasil Uji Wald dan Uji t disajikan di Tabel 9 dan 10.

TABEL 8. NILAI EFISIENSI RELATIF (ER)

Penduga Parameter	Fraksi Data Hilang (γ)	Efisiensi Relatif (ER)
$\hat{\beta}_0$	0.060	0.988
$\hat{\beta}_1$	0.017	0.997
$\hat{\beta}_2$	0.024	0.995
$\hat{\beta}_3$	0.103	0.980

Berdasarkan Tabel 8, hasil data imputasi sudah memenuhi toleransi batas nilai efisiensi relatif sebesar 5% ($ER > 0.95$). Pendugaan data hilang dengan menggunakan metode PMM dengan lima proses imputasi sudah layak untuk menduga data hilang pada *Automobile dataset*. Hal tersebut disebabkan nilai fraksi terbesar hanya 0.10.

TABEL 9. NILAI UJI WALD

W_1	$F_{(0.05;3;165,5)}$	Kesimpulan
61.16	0.117	Signifikan

TABEL 10. NILAI UJI T

Penduga parameter	t_{hitung}	df	t_{tabel}	Kesimpulan
$\hat{\beta}_0$	6.594	148.800	1.976	Signfikan
$\hat{\beta}_1$	4.384	169.800	1.974	Signfikan
$\hat{\beta}_2$	-11.511	167.700	1.974	Signfikan
$\hat{\beta}_3$	-2.026	119.500	1.980	Signfikan

Berdasarkan hasil Tabel 9 dan 10, semua variabel bebas berpengaruh signifikan terhadap variabel terikat, dalam hal ini *City mpg* dipengaruhi oleh *Compression ratio*, *Bore*, dan *Stroke*.

KESIMPULAN DAN SARAN

Kesimpulan

Berdasarkan hasil dan pembahasan tersebut, imputasi ganda dengan metode PMM dapat digunakan untuk menduga data hilang pada variabel bebas ($X_1, X_2, \cdots, X_n$) di model regresi linear apabila data yang hilang berpola monoton. PMM memiliki berbagai tahapan, yaitu tahap imputasi, analisis, dan kombinasi. Proses imputasi pada metode PMM dilakukan melalui pengisian data yang hilang dengan data yang berhasil diobservasi. Oleh karena itu, hasil imputasi dengan metode PMM selalu berada pada rentang data yang berhasil diobservasi.

Tingkat keakuratan imputasi ganda dengan metode PMM untuk menduga data hilang pada model regresi linear dapat diketahui berdasarkan nilai efisiensi relatif. Nilai efisiensi semakin baik apabila semakin mendekati satu. Hasil dan pembahasan dari skripsi ini menunjukkan bahwa imputasi sebanyak lima kali (m=5) hanya layak digunakan apabila nilai fraksi data hilang (γ) kurang dari atau sama dengan 0.25 pada batas toleransi 5% dari nilai efisiensi relatif. Selain nilai efisiensi relatif, hasil Uji Wald dan Uji t juga menentukan kelayakan suatu model regresi yang diperoleh dari hasil imputasi ganda dengan metode PMM.

Contoh penyelesaian kasus data hilang menggunakan imputasi ganda dengan metode PMM pada model regresi linear diterapkan pada *Automobile dataset*. Besarnya persentase data yang hilang adalah 2% dengan pola monoton yang tersebar pada Variabel *Bore* (X_2) dan *Stroke* (X_3). Hasil data imputasi memenuhi batas toleransi dari nilai efisiensi relatif sebesar 5%. Model regresi yang diperoleh menyatakan bahwa peningkatan satu persen dari *Compression ratio* akan meningkatkan nilai *City mpg* sebesar 2.113 satuan, sedangkan peningkatan satu inci dari *Bore* dan *Stroke*, akan

menurunkan nilai *City mpg* masing-masing sebesar -13.966 dan -2.642 satuan. Model tersebut layak digunakan berdasarkan hasil uji wald atau pun uji t dengan taraf signifikansi sebesar 5% ($\alpha = 5\%$). Dengan demikian, *Compression ratio*, *Bore*, dan *Stroke* memiiki pengaruh yang signifikan terhadap *City mpg*. Walaupun demikian, hanya *Compression ratio* yang berpengaruh positif terhadap *City mpg*.

Saran

Hal yang disarankan oleh penulis untuk peneliti selanjutnya adalah menguji efektivitas metode PMM apabila data yang hilang berada di variabel bebas dan terikat. Selain itu, disarankan juga untuk melakukan perbandingan tingkat akurasi metode imputasi ganda dengan PMM terhadap metode maksimum *likelihood* (algoritma EM) ketika menduga data hilang pada model regresi linear.

UCAPAN TERIMA KASIH

Terima kasih kepada Ibu Dra. Widyanti Rahayu, M.Si dan Ibu Vera Maya Santi, M.Si atas saran dan ketersediaan waktu dalam membimbing penulis, sehingga tulisan ini berhasil selesai dengan tepat waktu.

REFERENSI

Ahmad, dkk. 2014. *Optimization of an Internal Combustion Engine`s Efficiency for Fuel Conservation & Green Environment*. IEEE Paper: ICESP-2014-A-0086. DOI: 10.1109/ICESP.2014.7347006, hh. 1-9.

Buuren, S. V. 2012. *Flexible Imputation of Missing Data*. USA: CRC Press.

Little, R.J.A dan Rubin, D.B. 2002. *Statistical Analysis With Missing Data Second edition*. Canada: John Wiley & Sons, Inc.

Marshall, A., Altman, D.G., Holder,L.R., & Royston, P. 2009. *Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines*. BMC Medical Research Methodology Vol.9 No.57, hh. 205-214.

Moris, T. P., White, I.R., & Royston, P. 2014. *Tuning Multiple Imputation by Predictive Mean Matching and Local Residual Draws*. BMC Medical research Methodology Vol.14, hh. 75-87.

Schlimmer, J.C. 1987. *UCI Automobile Dataset*. [www.archive.ics.uci.edu/ml/datasets/automobile]. Pittsburgh, US: Carnegie Mellon School of Computer Science.

White, I.R., Royston, P., & Wood, A.M. 2010. *Multiple imputation using chained equations: Issues and guidance for practice*. USA: John Wiley & Sons, Ltd. No. 30, hh. 377-399.

Yan, Xin dan Su, X. G. 2009. *Linear Regression Analysis Theory and Computing*. Singapore: World Scientific Publishing Co. Pte. Ltd.