ANALISIS REGRESI SEMIPARAMETRIK PADA KASUS HILANGNYA RESPON

Irma Yahya[1)], I Nyoman Budiantara[2)], dan Kartika Fitriasari[2)]
[1)]Jurusan Matematika FMIPA, Universitas Haluoleo Kendari
[2)]Jurusan Statistika FMIPA, ITS Sukolilo Surabaya

Abstract. In the specific cases of experiment, not all data (response) may be available, which is called missing response cases. It's appear for various reasons. For the existing problem, inference statistics cannot be applied directly. The aim of this research is to consider about certain method to impute the missing response which is related to semiparametric regression, as a goodness of fit measurement of the used method, suppose an estimator $\hat{\theta}$ which is compared to the mean of complete response, then consider asymptotic distribution, consistency and efficiency of parametrics component estimator.
By using Kernel approximation, the resulted of nonparametrics estimator and by least square method, the resulted parametric component .The application to minimum temperature's data in 56 cities at USA, estimator value of $\hat{\theta}$ for several confidence interval tend to be similar to the mean value of complete response.

Keywords: Asymptotic, Kernel Estimator, Missing Response, Semiparametric Regression.

1. PENDAHULUAN

Berbicara tentang inferensi statistik, dimana teori probabilitas digunakan sebagai pondasinya atau dasarnya, sama halnya berbicara masalah estimasi, baik estimasi interval maupun estimasi titik dan masalah pengujian hipotesis. Ketika melakukan inferensi statistik dibutuhkan data yang lengkap. Namun tak dapat dipungkiri bahwa dalam suatu penelitian dengan berbagai alasan sering terjadi kehilangan informasi untuk mendapatkan data (respon) lengkap yang dibutuhkan, yang biasanya disebut sebagai kasus hilangnya (*missing*) respon. Misalnya karena ketidaksediaan dari unit-unit sampel untuk memberikan informasi atau karena adanya faktor-faktor yang tidak terkontrol. Untuk mengatasi masalah kehilangan respon tersebut, hal yang biasanya dilakukan yaitu dengan cara membuang nilai variabel-variabel prediktor yang bersesuaian dengan nilai respon yang hilang, tetapi hal ini tidak selamanya dapat dilakukan ketika kontribusi dari nilai variabel-variabel prediktor itu sangat dibutuhkan, atau dengan cara mengganti setiap respon yang hilang dengan suatu nilai yang wajar kemudian dilakukan analisis statistik berdasarkan data yang lengkap, namun hal ini juga akan mengakibatkan inferensi statistik dengan bias yang besar.

Dalam beberapa tahun terakhir ini telah banyak peneliti yang membahas tentang isu di atas dengan berbagai metode diantaranya yang berhubungan dengan regresi linier [11], [3], metode ratio [4], [7] mengawali metode kernel untuk missing respon, menggunakan estimasi regresi nonparametrik untuk mengestimasi respon yang hilang dengan asumsi MAR, [8] menggunakan densitas kernel yang dikombinasikan dengan nonparametrik *bootstrap*, Efron (1994) dengan pendekatan Bayesian *bootstrap*, [2] dengan pendekatan kernel untuk nonparametrik, pendekatan regresi multivariat [6], pendekatan *Likelihood* [4].

Dalam tulisan ini akan dibahas suatu metode untuk mengganti respon yang hilang didasarkan pada persamaan regresi semiparametrik:

$$Y_i = X_i^T \beta + g(T_i) + \varepsilon_i \ , \ \text{i} = 1,2,\ldots,\text{n} \quad (1.1),$$

dimana Y_i adalah variabel respon, X_i *dan* T_i adalah variabel prediktor, $g(.)$ adalah fungsi yang tidak diketahui dan ε_i adalah error yang independen dengan

mean nol dan varians σ^2. Diasumsikan Y_i hilang secara acak (*missing at random*, *MAR*) Untuk mengukur kebaikan metode yang digunakan, diberikan suatu ukuran [10] yang didefinisikan:

$$\theta = \frac{1}{n}\sum_{i=1}^{n}\left[\delta_i Y_i + (1-\delta_i)(X_i^T\beta + g(T_i))\right]. \quad (1.2)$$

Sebagai ukuran kebaikan dari metode ini, yaitu bahwa nilai $\hat{\theta}$ akan mendekati nilai rata-rata dan nilai estimasi kurva respon lengkap.

2. ESTIMASI FUNGSI $g(.)$ DAN PARAMETER β

Jika persamaan (1.1) dihubungkan dengan kasus hilangnya respon dan β diketahui sebagai parameter yang benar maka untuk mengestimasi fungsi $g_1(t)$ dan $g_2(t)$, dilakukan dengan langkah awal yaitu persamaan (1.1) dikalikan dengan suatu indikator δ_i, dimana δ_i=0 jika Y_i hilang (*missing*) dan δ_i=1 jika Y_i tidak hilang, sehingga persamaan (1.1) menjadi:

$$\delta_i Y_i = \delta_i X_i^T\beta + \delta_i g(T_i) + \delta_i \boldsymbol{\varepsilon_i} \quad (2.1)$$

Selanjutnya persamaan (2.1) diekspektasikan dengan syarat (T=t) maka diperoleh

$$E(\delta_i Y_i | T=t) = E(\delta_i X_i^T | T=t)\beta + E(\delta_i | T=t)g(t)$$

sehingga:

$$g(t) = \frac{E(\delta_i Y_i | T_i=t)}{E(\delta_i | T_i=t)} - \frac{E(\delta_i X_i^T | T_i=t)}{E(\delta_i | T_i=t)}\beta. \quad (2.2)$$

Persamaan (2.2) dapat ditulis sebagai berikut:

$$g(\mathbf{t_i}) = g_2(t_i) - g_1^T t_i)\beta, \quad (2.3)$$

dimana

$$g_2(t) = \frac{E(\delta_i Y_i | T_i=t)}{E(\delta_i | T_i=t)}, \quad g_1(t) = \frac{E(\delta_i X_i^T | T_i=t)}{E(\delta_i | T_i=t)}$$

Dengan menggunakan pendekatan fungsi kernel maka akan diperoleh estimator dari $g_1(t)$, $g_2(t)$, *dan* $g(t)$ masing-masing sebagai berikut

$$\hat{g}_1(t) = \frac{n^{-1}\sum_{i=1}^{n}\delta_i K_{h_1}(t-T_i)X_i}{n^{-1}\sum_{i=1}^{n}\delta_i K_{h_1}(t-T_i)} = \sum_{i=1}^{n}\delta_i W_{h_i}(t_i)X_i \quad (2.4)$$

$$\hat{g}_2(t) = \frac{n^{-1}\sum_{i=1}^{n}\delta_i K_{h_1}(t-T_i)Y_i}{n^{-1}\sum_{i=1}^{n}\delta_i K_{h_1}(t-T_i)} = \sum_{i=1}^{n}\delta_i W_{h_i}(t)Y_i \quad (2.5)$$

dimana $W_{ni}(t) = \dfrac{K(\frac{t-t_i}{h})}{\sum_{i=1}^{n}\delta_i K\left(\frac{t-t_i}{h}\right)}$.

Setelah estimator-estimator dari bagian nonparametrik diperoleh, selanjutkan ditentukan estimator parametrik yaitu $\hat{\beta}$, Untuk mengetimasi estimator ini digunakan metode kuadrat terkecil dan dengan meminimumkan

$$Q = \sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}\left[(Y_i - \hat{g}_{2n}(t_i)) - (X_i - \hat{g}_{1n}(t_i))^T\beta\right]^2, \quad (2.6)$$

maka diperoleh

$$\hat{\beta}_n = \left[\sum_{i=1}^{n}\delta_i(X_i - \hat{g}_{1n}(t_i))(X_i - \hat{g}_{1n}(t_i))^T\right]^{-1} \times \sum_{i=1}^{n}\delta_i[(X_i - \hat{g}_{1n}(t_i))(Y_i - \hat{g}_{2n}(t_i))]. \quad (2.7)$$

Estimator-estimator yang telah diperoleh, kemudian disubtitusikan kedalam persamaan (2.3) untuk menentukan estimator dari fungsi $g(.)$.

Sebagai ukuran kebaikan dari metode yang digunakan pada regresi semiparametrik pada kasus hilangnya respon, digunakan suatu ukuran kebaikan seperti pada persamaan (1.2) di atas, dengan memsubtitusikan estimator-estimator parametrik dan nonparametrik yang diperoleh.

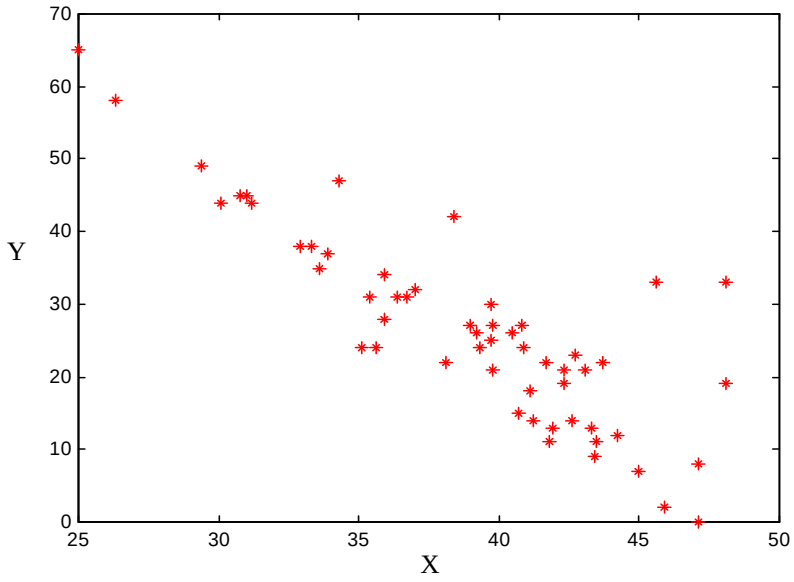
3. APLIKASI DATA

Untuk aplikasi digunakan rata-rata suhu minimum bulan Januari di 56 kota Amerika Serikat. Ingin diketahui bagaimana pengaruh letak suatu kota berdasarkan derajat bujur (longitude) dan derajat lintang (latitude) terhadap suhu rata-rata minimum.

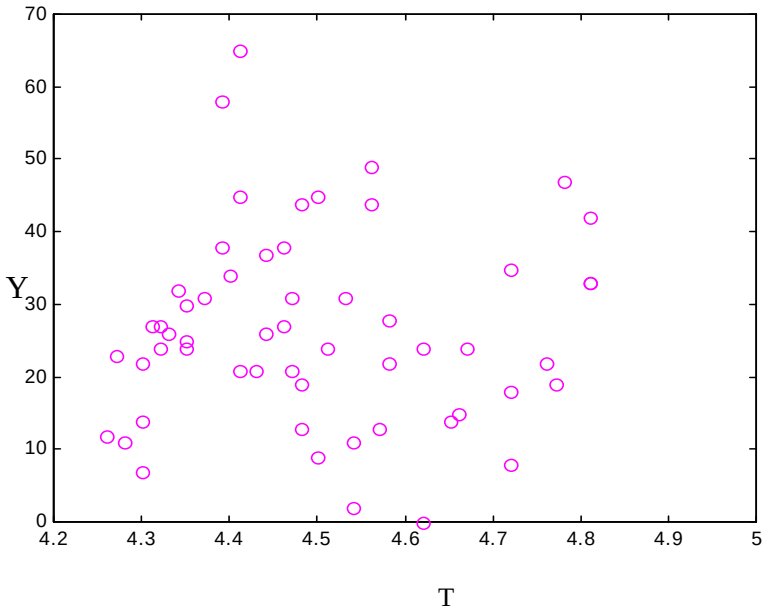
Sebagai langkah awal yaitu menentukan variabel-variabel prediktor yang mana sebagai variabel parametrik dan variabel nonparametrik. Salah satu cara untuk melihat hal tersebut yaitu dengan melihat plot antara masing-masing variabel prediktor dengan varibel respon, jika plot antara variabel prediktor dengan variabel respon mengarah ke suatu bentuk kurva tertentu maka variabel prediktor tersebut merupakan varibel parametrik sedangkan jika plot tersebut tidak jelas bentuk kurvanya maka variabel prediktor tersebut adalah variabel nonparametrik.

Berdasarkan Gambar (3.1) terlihat bahwa antara variabel X (lintang) dan variabel Y (suhu minimum), jelas plotnya mengarah ke suatu bentuk kurva tertentu sehingga variabel X ditetapkan sebagai varibel parametrik sedangkan dari Gambar (3.2) plot antara variabel T (bujur) dan variabel Y (suhu minimum) tidak mengarah ke suatu bentuk kurva tertentu sehingga variabel T ditetapkan sebagai variabel nonparametrik.

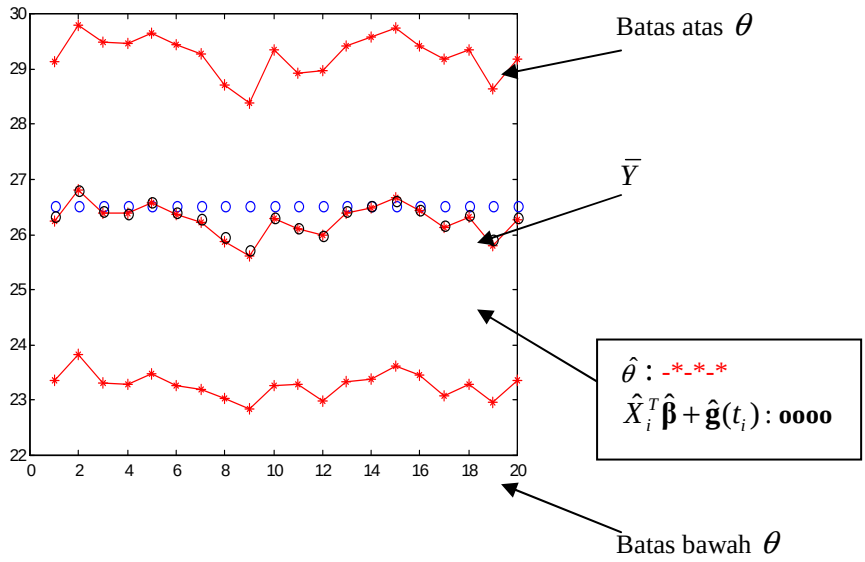
Pada proses hilangnya respon 5% dan 10 % dengan interval konvidensi 90% diperoleh nilai $\hat{\boldsymbol{\theta}}$ dan nilai estimasi kurva. Pada Tabel 3.1 di halaman lampiran, terlihat dengan jelas nilai-nilai $\hat{\boldsymbol{\theta}}$ hampir sama dengan rata-rata respon lengkap yaitu **26,5179** .



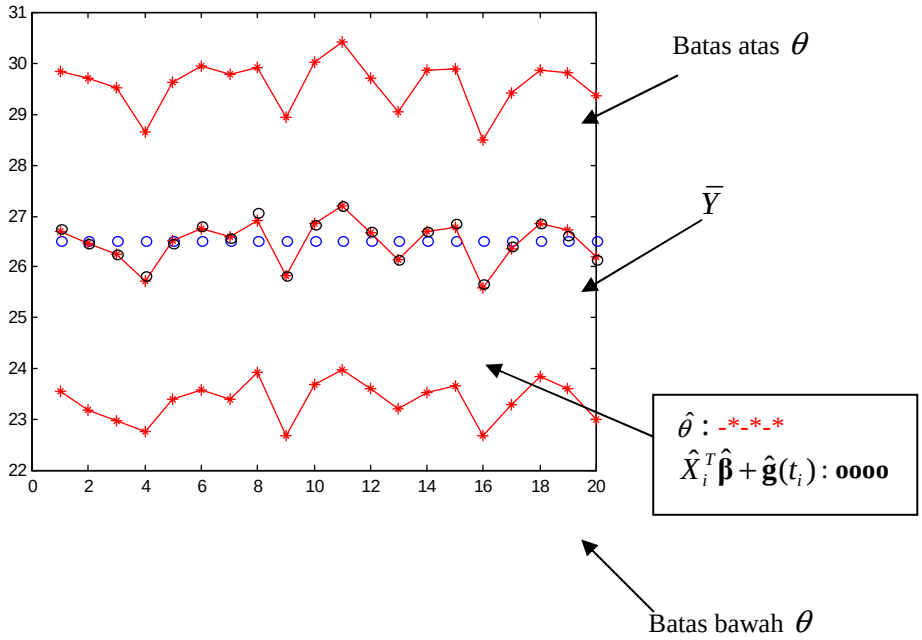
Gambar 3.1. Plot antara Suhu Minimum (Y) dan Garis Lintang(X)



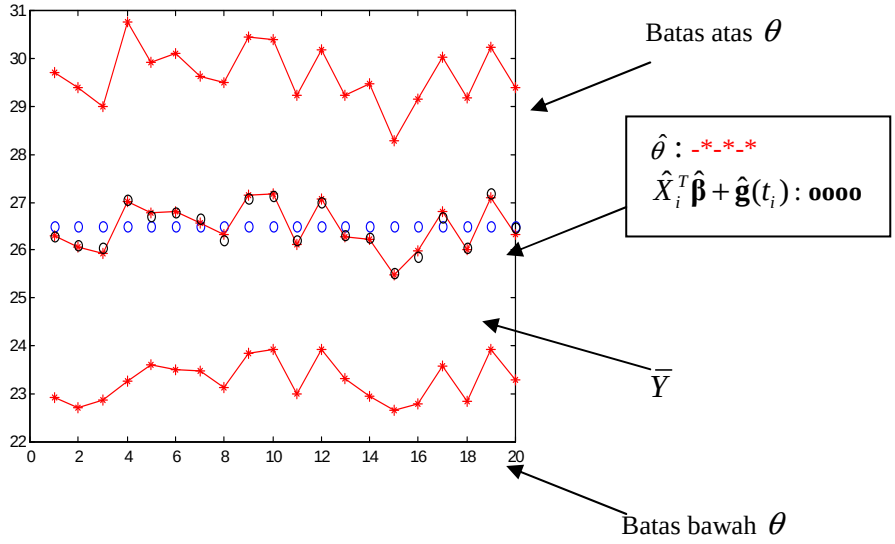
Gambar 3.2. Plot antara Suhu Minimum (Y) dan Garis Bujur (T)



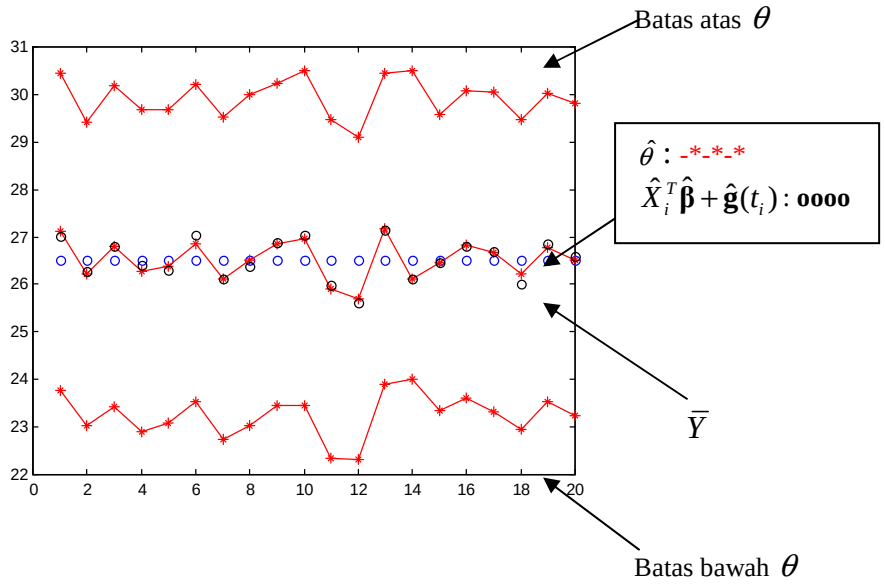
Gambar 3.3. Plot Nilai $\hat{\boldsymbol{\theta}}$, Interval Konfidensi 90 % untuk $\boldsymbol{\theta}$ dan Estimasi Kurva Regresi Hilangnya Respon 5%.



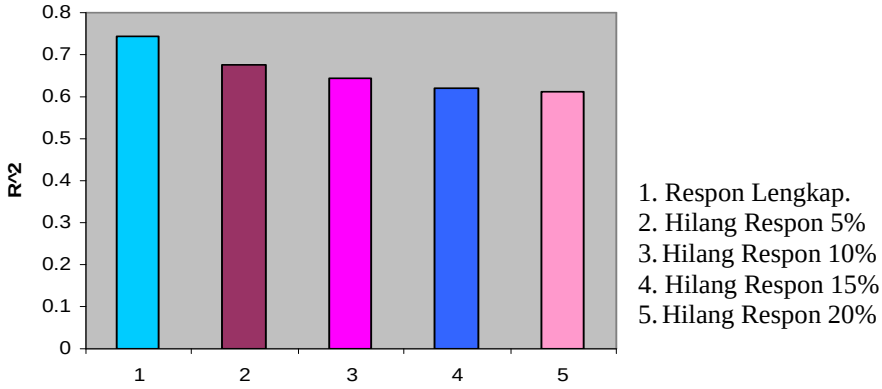
Gambar 3.4. Plot Nilai $\hat{\boldsymbol{\theta}}$, Interval Konfidensi 90 % untuk $\boldsymbol{\theta}$ dan Estimasi Kurva Regresi Hilangnya Respon 10%.



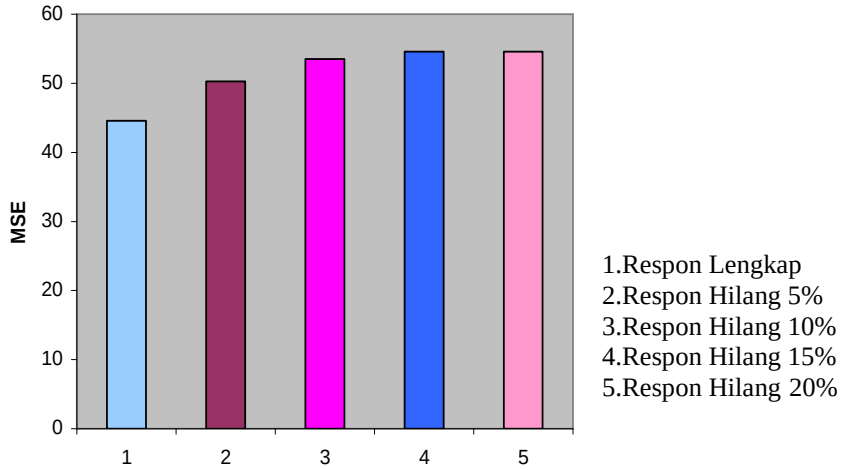
Gambar 3.3. Plot Nilai $\hat{\boldsymbol{\theta}}$, Interval Konfidensi 90 % untuk $\boldsymbol{\theta}$ dan Estimasi Kurva Regresi Hilangnya Respon 15%.



Gambar 3.4. Plot Nilai $\hat{\boldsymbol{\theta}}$, Interval Konfidensi 90 % untuk $\boldsymbol{\theta}$ dan Estimasi Kurva Regresi Hilangnya Respon 20%.



Gambar 3.5. Diagram Batang R^2 Respon Lengkap Dan Respon Hilang dengan Interval Konfidensi 90%.



Gambar 3.6. Diagram Batang MSE Respon Lengkap Dan Respon Hilang dengan Interval Konfidensi 90%.

Dari Gambar 3.3 dapat dilihat bahwa batas bawah θ terkecil 22,8340 dan terbesar adalah 23,8358 sedangkan batas atas terkecil 28,3859 dan terbesar 29,7925. Pada Gambar 3.4 diperoleh batas bawah terkecil 22,6811 dan terbesar 23,9884 serta batas atas terkecil 28,4958 dan yang terbesar 30,4136. Terlihat juga bahwa nilai-nilai estimasi kurva regresi dan nilai rata-rata respon lengkap berada diantara batas atas dan batas bawah tersebut.

Pada proses hilangnya respon 15% dan 20 % dengan interval konvidensi 90% diperoleh nilai $\hat{\boldsymbol{\theta}}$ dan nilai estimasi kurva. Pada Tabel 3.2 di halaman lampiran, terlihat bahwa nilai-nilai $\hat{\boldsymbol{\theta}}$ hampir sama dengan nilai rata-rata respon lengkapyaitu **26.5179**.

Dari Gambar 3.3 di atas dapat disimpulkan bahwa batas bawah θ terkecil 22,6688 dan terbesar adalah 23,9301 sedangkan batas atas terkecil 28,2899 dan terbesar 30,7545. Pada Gambar 3.4 diperoleh batas bawah terkecil 22,3042 dan terbesar 24,0130 serta batas atas terkecil 29,1067 dan yang terbesar 30,4998. Nilai estimasi kurva regresi dan nilai rata-rata respon lengkap berada diantara batas atas dan batas bawah.

Untuk nilai R^2 dan MSE dari hilangnya respon 5%, 10%, 15 dan 20% (Gambar 3.5 dan Gambar 3.6) di atas, nilai-nilai tersebut cenderung sama dengan nilai R^2 dan MSE dari respon lengkap, sehingga dapat disimpulkan bahwa proses penggantian respon yang hilang dengan mempergunakan metode ini adalah tidak merubah sifat dari respon lengkap atau metode ini cukup baik.

4. DAFTAR PUSTAKA

[1] Bartle,R.G. dan Sherbhet, D.R, (1982), *Introduction to Real Analysis*, John Wiley & Sons, Inc, New York.

[2] Cheng,P.E., (1994), *Nonparametric Estimation of Mean Functional with Data Missing at Random,* J. Amer. Statist. Assoc., **89**: 81-87.

[3] Healy,M.J.R. dan Westmacoot,M. (1996), *Missing Values in Experiments Analyzed on Automatic Computers,* J.App. Statist.

[4] Rao,J.N.K.,(1996), *On Variance Estimation with Impute Survey Data,* J. Amer. Statist Asso., **91**: 499-520.

[5] Rohatgi,V.K., (1976), *An Introduction to Probability Theory and Mathematical Statistics,* John Wiley & Sons, New York.

[6] Robins,J.dan Rotnizky,A., (1995), *Semiparametric Efficiency in Multivariate Regression Models with Missing Data,* Journal of the American Statistical Association, . **90**: 122-129.

[7] Titterington,D.M. dan Mill,G.M, (1983), *Kernel Based Density Estimates from Incomplete Data,* Journal of the Royal Statistical Society B, **45**:258-266.

[8] Titterington,D.M. dan Sedransk, J, (1989), *Imputation of Missing Values Using Density Estimation Statistics & Probability Letters,* **8**: 411-418.

[9] Wang,Q.H. dan Rao,J.N.K. (2002), *Empirical Likelihood for Linier Regression Model Under Imputation for Missing Respon,* The Canadian Journal of Statistics, **29**: 597-608.

[10] Wang, Q.H. dan Linton, O (2004), *Semiparametric Regression Analysis with Missing Response at Random,* Journal of the American Statistical Association, **99**: 334-345.

[11] Yates,F (1993), *The Analysis of Replicated Experiments Where Field Result are Incomplete,* J. Exp. Agric., **1**:129-142.

LAMPIRAN

Tabel Hasil 3.1. Nilai $\hat{\boldsymbol{\theta}}$, Interval Konfidensi 90% untuk $\boldsymbol{\theta}$ dan $\mathbf{X}_i^T\hat{\boldsymbol{\beta}} + \hat{\mathbf{g}}(\mathbf{t}_i)$

Hilangnya Respon (%)	Ulangan	$\hat{\theta}$	Interval Konfidensi 90% Untuk θ		$\mathbf{X}_i^T\hat{\boldsymbol{\beta}} + \hat{\mathbf{g}}(\mathbf{t}_i)$
			Batas Bawah	Batas Atas	
5	1	26.2491	23.3619	29.1364	26.3448
	2	26.8141	23.8358	29.7925	26.8141
	3	26.3981	23.3178	29.4784	26.4419
	4	26.3807	23.2916	29.4698	26.3758
	5	26.5733	23.4893	29.6572	26.5949
	6	26.3593	23.2710	29.4477	26.4050
	7	26.2295	23.1953	29.2638	26.2848
	8	25.8705	23.0206	28.7205	25.9597
	9	25.6099	22.8340	28.3859	25.7187
	10	26.3009	23.2559	29.3458	26.3080
	11	26.0945	23.2790	28.9101	26.1392
	12	25.9785	22.9911	28.9659	25.9956
	13	26.3791	23.3456	29.4127	26.4319
	14	26.4919	23.3955	29.5884	26.5251
	15	26.6791	23.6245	29.7336	26.6183
	16	26.4326	23.4555	29.4096	26.4626
	17	26.1310	23.0896	29.1725	26.1698
	18	26.3225	23.2926	29.3523	26.3528
	19	25.7946	22.9518	28.6373	25.9071
	20	26.2659	23.3581	29.1736	26.3151
10	1	26.7052	23.5640	29.8465	26.7602
	2	26.4492	23.1869	29.7115	26.4727
	3	26.2591	22.9881	29.5302	26.2513
	4	25.7123	22.7705	28.6541	25.8148
	5	26.5137	23.4106	29.6168	26.4716
	6	26.7639	23.5924	29.9354	26.7984
	7	26.5965	23.4109	29.7821	26.5746
	8	26.9183	23.9169	29.9198	27.0606
	9	25.8138	22.6835	28.9440	25.8154
	10	26.8594	23.6844	30.0344	26.8188
	11	27.2010	23.9884	30.4136	27.1883
	12	26.6650	23.6159	29.7141	26.7080
	13	26.1377	23.2209	29.0545	26.1394
	14	26.6989	23.5344	29.8634	26.7081
	15	26.7714	23.6602	29.8825	26.8660
	16	25.5885	22.6811	28.4958	25.6774
	17	26.3617	23.3004	29.4230	26.4039
	18	26.8593	23.8567	29.8619	26.8533
	19	26.7122	23.6117	29.8127	26.6282
	20	26.1926	23.0091	29.3761	26.1333

Tabel 3.2. Nilai $\hat{\theta}$, Interval Konfidensi 90 % untuk θ dan Estimasi Kurva Regresi

Hilangnya Respon (%)	Ulangan	$\hat{\theta}$	Interval Konfidensi 90% Untuk θ		$X_i^T\hat{\beta} + \hat{g}(t_i)$
			Batas Bawah	Batas Atas	
15	1	26.3114	22.9170	29.7058	26.2912
	2	26.0519	22.7181	29.3857	26.1133
	3	25.9368	22.8799	28.9936	26.0574
	4	27.0094	23.2641	30.7547	27.0560
	5	26.7645	23.6191	29.9098	26.7303
	6	26.8127	23.5101	30.1152	26.8108
	7	26.5593	23.4907	29.6278	26.6739
	8	26.3155	23.1401	29.4910	26.2220
	9	27.1389	23.8401	30.4378	27.0837
	10	27.1637	23.9289	30.3984	27.1456
	11	26.1178	23.0125	29.2231	26.2251
	12	27.0581	23.9301	30.1860	27.0170
	13	26.2776	23.3119	29.2432	26.3411
	14	26.2196	22.9594	29.4798	26.2871
	15	25.4794	22.6688	28.2899	25.5257
	16	25.9826	22.8035	29.1616	25.8780
	17	26.7973	23.5735	30.0211	26.6922
	18	26.0061	22.8422	29.1701	26.0735
	19	27.0829	23.9295	30.2362	27.2052
	20	26.3383	23.2851	29.3916	26.4934
20	1	27.1113	23.7766	30.4459	27.0194
	2	26.2142	23.0190	29.4094	26.2724
	3	26.8004	23.4289	30.1719	26.7950
	4	26.2763	22.8847	29.6678	26.4075
	5	26.3829	23.0923	29.6736	26.3140
	6	26.8638	23.5304	30.1972	27.0489
	7	26.1256	22.7418	29.5094	26.1190
	8	26.5065	23.0226	29.9904	26.3765
	9	26.8550	23.4624	30.2475	26.8933
	10	26.9743	23.4605	30.4881	27.0499
	11	25.9058	22.3553	29.4563	25.9749
	12	25.7055	22.3042	29.1067	25.6234
	13	27.1750	23.9090	30.4410	27.1497
	14	26.1270	24.0130	30.4998	26.1198
	15	26.4710	23.3586	29.5833	26.4618
	16	26.8346	23.5999	30.0694	26.7999
	17	26.6747	23.3109	30.0386	26.7075
	18	26.2147	22.9617	29.4677	26.0035
	19	26.7771	23.5304	30.0238	26.8652
	20	26.5236	23.2399	29.8072	26.5800