PENERAPAN *PARTIAL LEAST SQUARE* (PLS) UNTUK MEMPEROLEH MODEL BEBAS MULTIKOLINIER

SKRIPSI

Oleh:

Novika Herawaty
NIM 081810101018

JURUSAN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS JEMBER
2015

PENERAPAN *PARTIAL LEAST SQUARE* (PLS) UNTUK MEMPEROLEH MODEL BEBAS MULTIKOLINIER

SKRIPSI

diajukan guna melengkapi tugas akhir dan memenuhi salah satu syarat
untuk menyelesaikan Program Studi Matematika (S1)
dan mencapai gelar Sarjana Sains

oleh

Novika Herawaty
NIM 081810101018

JURUSAN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS JEMBER
2015

PERSEMBAHAN

Skripsi ini saya persembahkan untuk:

1. ayahanda Hery Susanto dan ibunda Sri Sundari tercinta yang senantiasa memberi doa, semangat, dan kasing sayang;
2. adek Dwi Cahyo Ali Syahbana tersayang yang senantiasa memberi semangat dan inspirasi;
3. suami Bambang Sulistiono terkasih yang senantiasa menemani dan memberi doa, dan semangat;
4. nenek Sumiani tercinta yang senantiasa memberi semangat dan inspirasi;
5. Dr. Alfian Futuhul Hadi, S. Si., M. Si selaku Dosen Pembimbing Utama dan Dian Anggraeni, S. Si., M. Si selaku Dosen Pembimbing Anggota yang telah memberikan bantuan dan bimbingan secara intensif untuk penyempurnaan skripsi ini;
6. seluruh guru dan dosen sejak taman kanak-kanak hingga perguruan tinggi yang telah memberi banyak ilmu dan membimbing dengan tulus;
7. almamater Jurusan Matematika FMIPA Universitas Jember.

MOTTO

"Bersikaplah kukuh seperti batu karang yang tidak putus-putusnya dipukul ombak. Ia tidak saja tetap berdiri kukuh, bahkan ia menenteramkan amarah ombak dan gelombang itu"

(Marcus Aurelius) *)

"Hiduplah seperti pohon kayu yang lebat buahnya; hidup di tepi jalan dan dilempari orang dengan batu, tetapi dibalas dengan buah"

(Abu Bakar Sibli) *)

---

*) http://www.maribelajarbk.web.id/2015/03/contoh-motto-terbaru-dalam-skripsi.html.

PERNYATAAN

Saya yang bertanda tangan di bawah ini:

nama : Novika Herawaty

NIM : 081810101018

menyatakan dengan sesungguhnya bahwa karya ilmiah yang berjudul "Penerapan *Partial Least Square* (PLS) Untuk Memperoleh Model Bebas Multikolinier" adalah benar-benar hasil karya sendiri, kecuali kutipan yang sudah saya sebutkan sumbernya, belum pernah diajukan dalam institusi manapun dan juga bukan karya jiplakan. Saya bertanggung jawab atas keabsahan dan kebenaran isinya sesuai dengan sikap ilmiah yang harus dijunjung tinggi.

Demikian pernyataan ini saya buat dengan sebenarnya, tanpa ada tekanan dan paksaan dari pihak manapun serta bersedia mendapat sanksi akademik jika ternyata di kemudian hari pernyataan ini tidak benar.

Jember, November 2015

Yang menyatakan,

Novika Herawaty

NIM 081810101018

SKRIPSI

PENERAPAN *PARTIAL LEAST SQUARE* (PLS) UNTUK MEMPEROLEH MODEL BEBAS MULTIKOLINIER

oleh

Novika Herawaty
NIM 081810101018

Pembimbing

Dosen Pembimbing Utama : Dr. Alfian Futuhul Hadi, S. Si., M. Si

Dosen Pembimbing Anggota : Dian Anggraeni, S. Si., M. Si

PENGESAHAN

Skripsi berjudul "Penerapan *Partial Least Square* (PLS) Untuk Memperoleh Model Bebas Multikolinier" telah diuji dan disahkan pada:

hari, tanggal :

tempat : Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Jember.

Tim Penguji:

Ketua,	Sekretaris,
Dr. Alfian Futuhul Hadi, S. Si., M. Si NIP. 197407192000121001	Dian Anggraeni, S. Si., M. Si NIP. 198202162006042002
Penguji I,	Penguji II,
Prof. Drs. I Made Tirta, M. Sc, Ph. D NIP. 195912201985031002	Kusbudiono, S. Si., M. Sc NIP. 197704302005011001

Mengesahkan

Dekan,

Prof. Drs. Kusno, DEA, Ph.D.
NIP. 196101081986021001

RINGKASAN

Penerapan *Partial Least Square* (PLS) Untuk Memperoleh Model Bebas Multikolinier; Novika Herawaty; 2015; 38 halaman; Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Jember.

Regresi linier merupakan metode statistika untuk membentuk hubungan antara variabel tak bebas (Y) dengan variabel bebas (X). Ada beberapa asumsi yang harus dipenuhi untuk membentuk model regresi linier, diantaranya adalah memiliki galat normal, tidak mengandung multikolinieritas, tidak terdapat autokorelasi, dan varian residual yang konstan (homoskedatisitas). Namun pada kenyataannya tidak semua data dapat memenuhi asumsi regresi linier sehingga regresi linier tidak dapat digunakan.

Salah satu uji asumsi regresi linier yang terkadang tidak dapat terpenuhi adalah multikolinieritas. Dimana kasus ini terjadi ketika adanya hubungan linier yang kuat diantara beberapa variabel bebas dalam suatu model regresi linier. Model regresi yang baik seharusnya memiliki variabel-variabel bebas yang tidak berkorelasi. *Partial Least Square* (PLS) salah satu cara yang dapat digunakan untuk mengatasi multikolinieritas.

PLS merupakan tekhnik analisis yang mengkombinasikan analisis komponen utama dengan regresi linier berganda. Tujuan dari PLS yaitu membentuk komponen yang dapat menangkap informasi dari variabel bebas untuk menduga variabel tak bebas. Penelitian ini bertujuan untuk mengetahui penerapan PLS untuk mendapatkan model yang bebas dari multikolinieritas.

Data yang digunakan dalam penelitian ini adalah data primer yang diperoleh dari data pengecekan kesehatan secara berskala terhadap siswa-siswi TK Pertiwi Jember tahun 2014-2015. Data meliputi umur siswa per Juli 2014 (4-7 tahun), tinggi badan, berat badan, lingkar kepala, dan lingkar lengan. Hasilnya metode PLS dapat

mengatasi multikolinieritas dan didapatkan model dengan 3 komponen berdasarkan nilai R^2 sebesar 0,7007 dan nilai MSE sebesar 14,76. Selanjutnya untuk dapat memberikan informasi pada koefisien penduga dari masing-masing variabel maka dilakukan transformasi dari model PLS ke model regresi linier. Model linier yang diperoleh dari model PLS dengan 3 komponen yaitu Y = 89,0029865 + 0,05746519LK – 1,2455682LL + 3,67830375UMUR + 1,10311800BB.

PRAKATA

Segala puji syukur kami panjatkan ke hadirat Allah SWT yang telah melimpahkan segala berkat, rahmat, serta hidayah-Nya, sehingga penulis dapat menyelesaikan skripsi yang berjudul "Penerapan *Partial Least Square* (PLS) Untuk Memperoleh Bebas Multikolinier". Skripsi ini disusun untuk memenuhi salah satu syarat dalam menyelesaikan pendidikan strata 1 (S1) di Jurusan Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Jember.

Penyusunan skripsi ini tidak lepas dari bantuan berbagai pihak, baik bantuan secara langsung maupun tidak langsung. Oleh karena itu, penulis menyampaikan terima kasih kepada:

1. ibu tercinta Sri Sundari yang telah sabar menanti selesainya skripsi ini;
2. ayah tersayang Hery Susanto yang selalu memberi pengertiannya;
3. adik terhebat Dwi Cahyo Ali Syahbana yang menjadi pacuan selesainya skripsi ini dan Nenek terkuat Sumiani yang selalu menjadi pendorong semangat;
4. suami terkasih Bambang Sulistiono yang setia menemani;
5. Bapak Dr. Alfian Futuhul Hadi, S. Si., M. Si yang telah memberi bimbingan dan petunjuk hingga setetes ilmu ini dapat terselesaikan;
6. Ibu Dian Anggraeni, S. Si, M. Si yang juga memberikan kasih sayang dan perhatian selama proses penyempurnaannya;
7. Prof. Drs. I Made Tirta M. Sc., Ph. D dan Kusbudiono, S. Si, M. Sc yang berkenan untuk menguji serta memberikan evaluasi agar selanjutnya buku ini dapat menjadi lebih bermanfaat;
8. mama Hj. Kurniati Ningsih dan papa Drs. H. Syamsul Arief yang telah sabar dan memberikan dukungannya dalam hal apapun;
9. Rafiantika Megahnia Prihandini, Masrurotullaily, dan Dimas Bagus C.W. yang selalu ada sebagai sahabat, saudara, dan guru sekaligus berkenan memberikan waktu dan ilmunya demi terselesaikannya skripsi ini;

10. Uma Fatimah Alkaff yang selalu memberikan doa dan kekuatannya agar selalu sabar menjalani semua;
11. Semua teman-teman yang bersedia bersama-sama mengejar mimpi dengan jalan masing-masing;
12. Setiap pribadi yang telah menjadi guru baik secara langsung maupun tak langsung.

Skripsi ini juga tidak lepas dari kekurangan dan kesalahan baik isi maupun susunannya. Oleh karena itu, penulis menerima segala kritik dan saran dari semua pihak demi kesempurnaan skripsi ini. Akhirnya penulis berharap semoga skripsi ini dapat memberi manfaat bagi pembaca dan kehidupan.

Jember, November 2015 Penulis

DAFTAR ISI

Halaman

DAFTAR GAMBAR

Halaman

DAFTAR TABEL

Halaman

DAFTAR LAMPIRAN

Halaman

BAB 1 PENDAHULUAN

1.1 Latar Belakang

Dalam analisis regresi, terkadang kita menjumpai kondisi terdapat korelasi antar variabel bebas atau yang biasa disebut dengan istilah multikolinier. Multikolinier menjadi masalah dalam analisis regresi. Multikolinier yang tinggi mengakibatkan tidak dapat melihat pengaruh variabel bebas terhadap variabel respon secara terpisah (Gujarati, 1992). Terdapat beberapa metode yang dapat mengatasi masalah multikolinier. Setiap metode mempunyai kekomplekan. Metode yang diusulkan untuk mengatasi masalah multikolinier salah satunya adalah *Partial Least Square* (PLS) (Yuliani, 2010)

Regresi PLS merupakan teknik yang mengeneralisasi dan mengkombinasikan analisis komponen utama dan regresi berganda (Abdi, 2006). PLS dapat digunakan untuk model yang mengandung sejumlah besar variabel bebas. PLS pertama kali diterapkan dalam bidang kemometrik (Geladi, 1992). Kemudian berkembang dan digunakan dalam bidang-bidang lain. PLS mereduksi dimensi variabel-variabel penjelas asal melalui pembentukan variabel-variabel laten dengan dimensi yang lebih kecil yang merupakan kombinasi linier dari variabel-variabel penjelas asal.

Menurut Ohyver (2013), dengan menggunakan gabungan antara transformasi logaritma natural dan PLS yang diterapkan pada data pengunjung hotel di Kendari, Sulawesi Utara dilakukan untuk mengatasi kasus multikolinier yang memenuhi asumsi normalitas sehingga diperoleh model yang bebas multikolinier dan *outlier*. Selain itu model yang diperoleh juga masih mempunyai nilai R^2 yang kecil sehingga variabel bebas tidak dapat memberikan pengaruh terhadap variabel terikat.

Berdasarkan latar belakang dan penelitian terdahulu di atas, maka penulis tertarik untuk melakukan penelitian yang berjudul **"Penerapan *Partial Least Square* (PLS) Untuk Memperoleh Model Bebas Multikolinier"**. Penelitian ini mengunakan data pemeriksaan kesehatan pada TK Pertiwi Jember tahun ajaran 2014-2015. Pemeriksaan kesehatan dilakukan tiga kali dalam setahun. Ini bertujuan untuk mengetahui tumbuh kembang anak pada usia 4-7 tahun. Pemeriksaan kesehatan dilakukan dengan mengetahui umur, tinggi badan, berat badan, lingkar kepala dan lingkar lengan pada anak.

1.2 Rumusan Masalah

Permasalahan yang akan dibahas dalam skripsi ini adalah bagaimana penerapan PLS untuk memperoleh model yang bebas dari multikolinier untuk data tinggi badan siswa–siswi TK Pertiwi Jember.

1.3 Tujuan Penelitian

Untuk mengetahui penerapan PLS dalam memperoleh model yang bebas dari multikolinier untuk data tinggi badan siswa-siswi TK Pertiwi Jember.

1.4 Manfaat Penelitian

Adapun manfaat yang dapat diambil dari penelitian ini adalah dapat menambah wawasan dan pengetahuan tentang PLS untuk memperoleh model yang bebas dari multikolinier untuk data tinggi badan siswa–siswi TK Pertiwi Jember.

BAB 2 TINJAUAN PUSTAKA

2.1 Regresi Linier

Model peramalan atau pendugaan pertama kali diperkenalkan oleh Sir Francis Galton (1822–1911), yang selanjutnya dinamakan regresi. Dalam penelitiannya, ia membandingkan tinggi badan anak laki- laki dengan tinggi badan ayahnya. Galton menunjukkan bahwa tinggi badan anak laki-laki dari ayah yang tinggi setelah beberapa generasi cenderung mundur (*regressed*) mendekati nilai tengah populasi. Dengan kata lain, anak laki-laki dari ayah yang badannya sangat tinggi cederung lebih pendek dari pada ayahnya, sedangkan anak laki-laki dari ayah yang badannya sangat pendek cenderung lebih tinggi dari ayahnya. Persamaan matematik yang memungkinkan untuk meramalkan nilai-nilai variabel tak bebas (Y) dari nilai-nilai satu atau lebih variabel bebas (X) disebut persamaan regresi.

Regresi linier adalah metode statistika yang digunakan untuk membentuk hubungan antara variabel tak bebas dengan variabel bebas. Karena antara variabel tak bebas dan variabel bebas memiliki hubungan, maka nilai variabel bebas dapat digunakan untuk menduga atau mearamal nilai variabel tak bebas. Dikatakan variabel bebas karena variabel ini nilainya tidak tergantung pada variabel lain. Sedangkan untuk variabel tak bebas nilainya bergantung pada variabel lain. Hubungan yang terbentuk dapat melibatkan satu atau lebih variabel tak bebas dengan satu atau lebih variabel bebas. Secara umum model regresi linier dapat ditulis dalam bentuk sebagai berikut:

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \cdots + \beta_p \mathrm{X}_p + \varepsilon \qquad (2.1)$$

Dengan:

Y = variabel tak bebas

Xj = variabel bebas untuk $j = 1, 2, \ldots, p$

βi = koefsien variabel X, i = 1,2,…, p

ε = galat (*error*)

Misalkan terdapat n buah pengamatan, maka persamaan (2.1) dapat ditulis menjadi:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon; \quad i = 1, 2, \dots, n \qquad (2.2)$$

Jika persamaan (2.2) ditulis dalam matriks, maka diperoleh persamaan seperti pada persamaan (2.3) dengan mendefinisikan matriks–matriksnya sebagai berikut:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \qquad (2.3)$$

(Tirta, 2009).

Y adalah vektor pengamatan berordo $n \times 1$, **X** adalah matriks variabel bebas berordo $n \times (p+1)$, $\boldsymbol{\beta}$ adalah vektor koefisien regresi berordo $(p+1) \times 1$ dan $\boldsymbol{\varepsilon}$ adalah vektor variabel bebas berordo $n \times 1$.

Beberapa asumsi-asumsi yang digunakan untuk model regresi adalah:

a. $\boldsymbol{\varepsilon_i}$ dan $\boldsymbol{\varepsilon_j}$ tidak berkorelasi sehingga:

$$cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j, i = 1,2, \dots, n, \ j = 1,2, \dots, p$$

b. $\boldsymbol{\varepsilon_i}$ merupakan suatu variabel bebas dengan *mean* sama dengan nol dan varians konstan yang dapat dinyatakan sebagai berikut:

$$\varepsilon_i \sim N(0, \sigma^2) \qquad E(\varepsilon) = 0 \qquad V(\varepsilon) = \sigma^2$$

c. parameter $\beta_1, \ \beta_2, \ \dots, \beta_p$ berupa konstanta:

Berdasarkan asumsi, maka kita dapat menuliskan persamaan dalam bentuk nilai harapan sebagai berikut:

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$ (Utomo, 2010).

Analisis regresi merupakan salah satu teknik analisis data dalam statistika yang seringkali digunakan untuk mengkaji hubungan antara beberapa variabel dan meramal suatu variabel, dengan terlebih dahulu menentukan satu variabel yang disebut dengan variabel tak bebas dan satu atau lebih variabel tak bebas. Untuk mengkaji hubungan atau pengaruh dua atau lebih variabel bebas terhadap variabel tak bebas, maka model regresi yang digunakan adalah model regresi linier berganda *(multiple linier regression model).*

Bentuk umum model regresi linier berganda dengan p variabel bebas adalah sebagai berikut:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

Dimana:

Y_i adalah variabel respon untuk pengamatan ke i, untuk $i = 1,2, \dots, n$

$\beta_0, \beta_1, \beta_2, \dots, \beta_{p-1}$ adalah parameter

$X_{i1}, X_{i2}, \dots, X_{i,p-1}$ adalah variabel bebas

ε_i merupakan residu atau error untuk pengamatan ke i yang diasumsikan berdistribusi normal yang saling bebas dan identik $(0, \sigma^2)$.

2.2 Normalitas

Dalam analisis regresi dan korelasi yang diasumsikan menyebar normal adalah residual e_i sehinga ada suatu pemikiran yang perlu di uji kenormalannya adalah residual, tetapi banyak juga yang melakukannya langsung terhadap data pengamatan, tepatnya terhadap peubah respon (peubah tak bebas Y). Keduanya sama saja karena berdasarkan sifat dari peubah acak yang menyebar normal, jika peubah tersebut menyebar normal maka kombinasi liniernya juga akan menyebar normal. Jadi jika residual menyebar normal maka Y juga menyebar normal karena Y adalah kombinasi linier dari residual e_i atau $Y_i = a + b X_i + e_i$.

Di samping itu, dalam melakukan uji koefisien regresi atau koefisien korelasi biasa digunakan sebaran t atau untuk pengujian secara simultan digunakan sebaran f. Kedua sebaran tersebut diturunkan/berasal dari sebaran normal. Atau untuk lebih jelasnya sebaran t dibangkitkan dari rasio dua peubah acak yang menyebar normal baku dan sebaran khi-kuadrat, sedangkan sebaran f dibangkitkan dari rasio dua peubah acak yang masing-masing menyebar khi-kuadrat. Sebaran khi-kuadrat sendiri berasal dari sebaran normal baku (sebaran normal baku jelas berasal dari sebaran normal).

Asumsi normalitas adalah asumsi residual yang berdistribusi normal. Asumsi ini harus terpenuhi untuk model regresi linier yang baik. Uji normalitas dilakukan pada nilai residual model regresi. Penyebab terjadinya kasus normalitas adalah:

a. Terdapat data residual dari model regresi yang memiliki nilai data yang berada jauh dari himpunan data atau data ekstrim (*outliers*), sehingga penyebaran datanya menjadi non-normal.
b. Terdapat kondisi alami dari data yang pada dasarnya tidak berdistribusi normal atau berdistribusi lain, seperti: distribusi binormal, multinormal, eksponensial, gamma, dll.

Berikut diberikan cara-cara mengidentifikasi adanya kasus normalitas:

a. Dilakukan pemeriksaan dengan metode grafik, yaitu pemeriksaan normalitas dengan output normal P-P plot atau Q-Q plot. Asumsi normalitas terpenuhi ketika pencaran data residual berada disekitar garis lurus melintang.
b. Dilakukan pengujian dengan metode formal, seperti: pengujian normalitas yang dilakukan melalui uji *Kolmogorov-Smirnov*, uji *Anderson-Darling*, uji *Shapiro-Wilk*, dan uji *Jarque-Bera* yang mana semua pengujian ini memiliki hipotesis interprestasi, yaitu:

   H_0 : residual berdistribusi normal

   H_1 : residual tidak berdistribusi normal

Asumsi normalitas terpenuhi ketika pengujian normalitas menghasilkan *P-value* lebih besar dari α dengan nilai α ditentukan sebesar 1%, 5%, atau 10%.

Banyak jenis metode yang dapat digunakan untuk menguji normalitas dari sejumlah distribusi data, salah satunya adalah dengan menggunakan metode *Shapiro-Wilk*. Metode *Shapiro-Wilk* adalah sebuah metode untuk mengolah sekumpulan data besar yang belum diolah dalam tabel distribusi frekuensi dengan cara mengurutkan data dan membaginya menjadi dua kelompok untuk dikonversikan dalam *Shapiro-*

wilk, dilanjutkan dengan transformasi Z untuk mendapatkan luas kurva normalnya. Transformasi dalam nilai Z dapat dilakukan juga tidak. Lihatlah rumus umum untuk metode *Shapiro wilk* dibawah ini:

$$T_3 = {}^1/_D \left[\sum_{i=1}^{k} a_1 \left(X_{n-i+1} - X_i\right)\right]^2 \tag{2.4}$$

Dengan keterangan:

a_1 : koefisien test dari *Shapiro-wilk*

D : didefinisikan dengan rumus

$$D = \sum_{i=1}^{n} (X_i - X_{bar})$$

X_{n-i+1} : nilai ke n – i + 1 pada data

X_i : nilai ke-I dalam data

X_{bar} : nilai rata-rata

$$G = b_n + c_n + ln[(T3 - d_n)/(1 - T_3)]$$

Dengan keterangan:

T_3 : didefinisikan dalam rumus *shapiro-wilk*

bn , cn , dn : konversi statistik *Shapiro wilk* pendekatan untuk distribusi normal

G : mirip dengan nilai Z pada distribusi normal.

Adapun syarat dalam menggunakan metode *Shapiro wilk* ini dalam uji normalitas yakni:

a. Data harus berskala interval atau ratio bersifat kuantitatif atau berupa nilai hitung
b. Data bersifat tunggal atau belum dikelompokkan dalam tabel distribusi frekuensi
c. Data diperoleh dari pengambilan sampel atau contoh secara acak atau random

Adapun signifikansi dari metode Shapiro wilk ini dalam uji normalitas adalah:

Siginifikansi: Jika nilai p > 5%, maka H_0 diterima dan H_1 ditolak. Jika nilai p < 5%, maka H_0 ditolak dan H_1 diterima. Jika digunakan rumus G, maka digunakan tabel distribusi normal. Signifikansi (sig) uji nilai T_3 Jika dibandingkan dengan nilai pada tabel *Shapiro wilk* untuk dibandingkan nilai probabilitas atau peluang (p) nya. Sehingga, signifikansinya memiliki konsep seperti signifikansi diatas.

2.3 Multikoliniearitas

Multikolinieritas adalah suatu kondisi dimana terjadi korelasi yang kuat diantara variabel-variabel bebas yang diikutsertakan dalam pembentukan model regresi linier. Jelas bahwa multikolinieritas adalah suatu kondisi yang menyalahi asumsi regresi linier. Tentu saja, multikolinieritas tidak mungkin terjadi apabila variabel bebas yang diikutsertakan hanya satu.Multikolinieritas mengakibatkan ragam dari koefisien regresi bertambah atau mengalami inflasi, untuk itulah faktor inflasi ragam VIF (*variance inflation factors)* merupakan salah satu indikator adanya multikolinieritas dalam variabel prediktor X_j.

VIF dari suatu variabel prediktor adalah ukuran kedekatan hubungan antara variabel prediktor X_j dengan variabel prediktor yang lain. VIF didapatkan dari meregresikan dengan variabel prediktor X_j menggunakan penduga OLS, di mana $j \neq j'$. Namun Boulesteix dan Stimmer (2005) menjelaskan bahwa ketika pengamatan (n) jauh lebih kecil ukurannya dibanding variabel prediktor (p), metode regresi linier dengan penduga OLS tidak dapat digunakan.

Hal ini disebabkan matriks kovarian $X^T X$ dengan ukuran $p \times p$ yang mempunyai pangkat lebih dari *n-1* adalah matriks singular. Oleh karena itu penduga bagi variabel respon tidak memberikan hasil yang akurat, akibatnya nilai VIF juga tidak bisa dihitung dengan akurat. Dodge (2008) mengatakan bahwa dalam matriks variabel prediktor X pada regresi linier berganda**,** kolinieritas atau multikolinieritas berarti satu kolom matriks X memiliki kombinasi linier dengan X kolom lainnya. Kombinasi linier antar kolom pada matriks variabel prediktor dapat diukur menggunakan matriks korelasi antar variabel prediktor.

Elemen-elemen dari matriks korelasi adalah nilai koefisien korelasi antara satu variabel prediktor dengan variabel prediktor yang lain. Misalnya akan dihitung korelasi antara variabel prediktor X_1 dan variabel prediktor X_2 maka persamaan yang digunakan adalah:

$$r_{12} = \frac{(\sum_{i=1}^{n} x_{1i}x_{2i})}{\sqrt{(\sum_{i=1}^{n} x_{1i}^2)(\sum_{i=1}^{n} x_{2i}^2)}}$$

Keterangan:

r_{12} : koefisien korelasi antara variabel X_1 dengan X_2

X_{1i} : $X_{1i} - \bar{X}_1$

X_{2i} : $X_{2i} - \bar{X}_2$

Untuk mengetahui apakah dua variabel saling berkorelasi, maka dilakukan pengujian koefisien korelasi dengan hipotesis sebagai berikut:

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

H_0 berarti hipotesis awal penelitian dan H_1 berarti hipotesis alternatif. *Rho* (ρ) pada hipotesis tersebut berarti parameter koefisien korelasi. Statistik uji yang digunakan adalah statistik uji t dengan persamaannya adalah:

$$t = \frac{r}{S_r} \tag{2.5}$$

dimana t adalah statistik uji yang didapatkan dari perhitungan (t_{hitung}), S_r adalah standart deviasi dari penduga r, dan r adalah penduga koefiisien korelasi yang didapatkan dari persamaan (2.5). Persamaan untuk menghitung S_r adalah:

$$S_r = \sqrt{\frac{1-r^2}{n-2}}$$

Di bawah H_0, statistik uji t mengikuti distribusi t dengan derajat bebas *n-2*. H_0 akan ditolak jika

$$|t| > {}^{t_a}/_{2,n-2}$$

a merupakan tingkat kepercayaan dan biasanya ditentukan sebesar 5%, ${}^{t_a}/_{2,n-2}$ merupakan nilai kritis untuk penolakan H_0. H_0 ditolak berarti dapat dikatakan bahwa dua variabel saling berkorelasi. Selain menggunakan statistik uji, penarikan kesimpulan suatu hipotesis juga dapat dilakukan dengan menggunakan nilai peluang (*p-value*). Nilai peluang ini dilambangkan dengan *p* dan didapatkan dari:

$$p = 2\,P\left(T > \frac{r}{S_r}\right) = 2P(T > t)$$

di mana T merupakan *studentized distribution* atau distribusi t. Untuk tingkat kepercayaan α, perbandingan *p* dengan α memungkinkan untuk membuat suatu kesimpulan bagi ditolak atau tidaknya suatu hipotesis nol yaitu jika $p < \alpha$ maka H_0 ditolak.

VIF merupakan salah satu cara yang dapat digunakan untuk mendeteksi gejala multikolinearitas pada analisis regresi. VIF tidak lain adalah mengukur keeratan hubungan antar variabel bebas. Cara menghitung VIF dengan mencari fungsi regresi linier dari masing-masing variabel bebas kemudian dihitung nilai R^2 pada model masing-masing variabel bebas. Rumus VIF yaitu:

$$\text{VIF} = 1 / (1 - R^2) \qquad (2.6)$$

Beberapa cara mengindentifikasi adanya multikoliniearitas pada model regresi, diantaranya adalah:

a. Jika nilai regresi menunjukkan nilai R^2 yang tinggi dan F statistik yang sangat signifikan (g*oodness of fit* terpenuhi), namun sebagian besar variabel bebas tidak signifikan pengaruhnya (t hitung kecil)
b. Terdapat korelasi yang tinggi ($R > 0.8$) antara satu pasang atau lebih variabel bebas dalam model
c. Mencari nilai *Condition Index (CI).* Condition indek yang bernilai lebih dari 30 mengindentifikasikan adanya multikolineritas
d. Dapat pula melihat indikasi multikolinearitas dengan *Tolerance Value* (TOL), *Eigenvalue*, dan yang paling umum digunakan adalah *Varians Inflation Factor* (VIF). nilai VIF > 10 mengindentifikasi adanya multikolinieritas.
e. Perubahan kecil sekalipun pada data akan menyebabkan perubahan signifikan pada variabel yang diamati.

f. Nilai koefisien variabel tidak sesuai dengan hipotesis, misalnya variabel yang seharusnya memiliki pengaruh positif (nilai koefisien positif), ditunjukkan dengan nilai negatif.

Beberapa dampak multikoliniearitas yang serius pada sebuah regresi, akan berdampak pada:

a. Varian besar (dari takasiran OLS)
b. Interval kepercayaan lebar (varian besar - Standar Error besar - interval kepercayaan lebar)
c. Uji t (t rasio) tidak signifikan, nilai t statistik menjadi lebih kecil sehingga variabel bebas tersebut menjadi tidak signifikan pengaruhnya. pengaruh lebih lanjutnya adalah bahwa koefisien regresi yang dihasilkan tidak mencerminkan nilai yang sebenarnya dimana sebagian koefisien cenderung *over estimate* dan yang lain *under estimate*
d. Terkadang taksiran koefisien yang didapat akan mempunyai nilai yang tidak sesuai dengan substansi, sehingga dapat menyesatkan interpretasi.

Hal-hal yang perlu dilakukan bila terjadi multikoliniearitas adalah:

a. Transformasi data (misalnya dengan logaritma natural)
b. Mengeluarkan variabel yang berkorelasi dalam model
c. Mencari data tambahan

2.4 Regresi *Partial Least Square* (PLS)

Menurut Abdi (2003), Regresi PLS adalah tekhnik analisis yang mengkombinasikan analisis komponen utama dengan regresi linier berganda. Teknik ini dapat digunakan untuk memprediksi variabel respon Y_i berdasarkan variabel prediktor X_i yang sangat banyak. Ketika banyaknya variabel prediktor lebih besar dibanding banyaknya pengamatan, matriks X^TX yang mempunyai pangkat lebih dari n-1 menjadi singular dan pendekatan regresi berganda dengan penduga OLS akan menjadi tidak akurat.

Metode regresi PLS mencari komponen dari X_i yang juga berkaitan dengan Y_i. Tujuan utama dari PLS adalah membentuk komponen yang dapat menangkap informasi dari peubah bebas untuk menduga peubah respon (Hoskuldsson dalam Garthwaite, 1994).

Regresi PLS bekerja dengan menguraikan variabel respon **y** dan variabel prediktor **X** dengan persamaan:

$$X = TP^T + E$$
$$y = Tq^T + f \qquad (2.7)$$

Dengan keterangan sebagai berikut:

y(n×m**)** = matriks variabel respon dengan ukuran n×m

T(n×h**) =** matriks variabel laten atau skor komponen PLS dengan ukuran n×h

q(m×h**) =** vektor faktor loading dari ***y*** dengan ukuran m×h

f(n×m) **=** vektor residual dari ***y*** dengan ukur n×m

X(n×p) = matriks variabel prediktor dengan ukuran n×p

P(p×h**) =** matriks faktor *loading* dari ***X*** dengan ukuran p×h

E(n×p) = matriks residual dari ***X*** dengan ukuran n×p

n = banyaknya pengamatan (*i* = 1,2, ... , n)

p = banyaknya variabel prediktor (*j* = 1,2, ... *p*)

m = banyaknya variabel respon dan dalam penelitian ini m=1

h = banyaknya variabel laten atau skor komponen

Persamaan (2.7) merupakan model regresi laten yaitu model regresi antara variabel respon **y** dengan variabel laten **T** .

Variabel laten **T** merupakan kombinasi linier dari matriks variabel prediktor **X** yaitu: $\boldsymbol{T} = \boldsymbol{XW}$

di mana **W** adalah matriks pembobot dengan ukuran $p \times h$, dengan . Kolom dari $a = 1,2, \dots, h$ matriks **T** dan **W** didefinisikan sebagai $w_a = (w_{1a}, \dots, w_{pa})$ dan $t_a = (t_{1a}, \dots, t_{na})$. Hubungan antara kolom dari matriks **T** dan kolom dari matriks **W** adalah:

$$t_1 = w_{11}X_1 + \cdots + w_{ph}X_p$$

$$t_h = w_{1h}X_1 + \cdots + w_{ph}X_p$$

di mana $\boldsymbol{w}_{ja}$ untuk j = 1,2, ...p dihitung berdasarkan persamaan:

$$w_a = max\left(cov(X_a w_a, y_a)\right)$$

dengan batasan:

${w_a}^T w_a = 1$ dan $cov(X_a w_a, X_{a'} w_{a'})$ = 0 untuk $a \neq a'$ dengan a = 1,2, .. , h,.

Perhitungan $\mathbf{w}_a$ dapat juga ditulis sebagai berikut :

$$w_{ja} = \frac{cov\,(y_a\,, x_{ja})}{\sqrt{\sum_{j=1}^{p} cov\,(y_a\,, x_{ja})^2}} \tag{2.8}$$

Pembilang pada persamaan (2.8) dalam bentuk matriks dapat juga ditulis sebagai $\mathbf{X}^\mathbf{T}\mathbf{y}$ dan penyebut pada persamaan (2.8) merupakan normalisasi dari matriks $\mathbf{X}^\mathbf{T}\,\mathbf{y}$.

Setelah didapatkan skor komponen melalui persamaan (2.7), hubungan antara elemen kolom matriks skor komponen PLS, $t_a = (t_{1a}, \dots, t_{na})^T$dan elemen dari kolom matriks faktor *loading* dari **X** yaitu $p_a = \left(p_{1a}, \dots, p_{pa}\right)^T$adalah:

$$X_{n1} = t_{n1}P_{11} + t_{n2}P_{12} + \cdots + t_{na}P_{1a} + e_{n1}$$

$$\dots = \qquad \dots. \tag{2.9}$$

$$X_{np} = t_{n1}P_{p1} + t_{n2}P_{p2} + \cdots + t_{na}P_{pa} + e_{np}$$

Perhitungan elemen matriks faktor *loading* **P** pada a = 1,2, ...,h dapat juga ditulis:

$$p_{ja} = \frac{cov\,(t_a\,, x_{ja})}{var(t_a)}$$

Untuk setiap $\boldsymbol{x_{ja}}$ pada a = 1,2, ...,h dihitung dengan menggunakan persamaan:

$$\boldsymbol{x}_{j(a)} = \boldsymbol{x}_{j(a-1)} - \boldsymbol{p}_{j1}\boldsymbol{t}_1 - \cdots - \boldsymbol{p}_{j(a-1)}\boldsymbol{t}_{(a-1)}$$

Perhitungan elemen faktor *loading* **q** pada a = 1,2, ...,h adalah:

$$q_a = \frac{cov\,(t_a\,, y_a)}{var(t_a)}$$

Jika sudah didapatkan skor komponen t sebanyak h komponen ($\mathbf{t}_h$) maka dilakukan regresi **y** terhadap $\mathbf{t}_h$ dan dapat dituliskan sebagai:

$$\boldsymbol{y} = \boldsymbol{q}_1\boldsymbol{t}_1 + \boldsymbol{q}_2\boldsymbol{t}_2 \ldots + \boldsymbol{q}_h\boldsymbol{t}_h \tag{2.10}$$

Jika $\mathbf{t}_h$ pada persamaan (2.10) disubstitusi dengan $\mathbf{t}_h$ pada persamaan (2.9) maka model regresi pada persamaan (2.10) ekuivalen dengan $\boldsymbol{y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ jika $\widehat{\boldsymbol{\beta}}$ disubstitusi menjadi $\widehat{\boldsymbol{\beta}} = \boldsymbol{W}\boldsymbol{q}'$. (Boulesteix dan Stimmer, 2005)

Algoritma yang digunakan dalam penelitian ini adalah algoritma PLS yang diperkenalkan oleh Wold (1983). Algoritma PLS merupakan algoritma yang digunakan untuk membangun model regresi PLS dengan kasus satu variabel respon. Langkah-langkah algoritma PLS dalam operasi matriks yang dijelaskan oleh Kondylis (2006) untuk $i = 1, \ldots, n$ dan $j = 1, \ldots, p$ serta $\boldsymbol{X}_{n\times p}$ dan $\boldsymbol{y}_{n\times 1}$ adalah sebagai berikut:

a. Matriks masukan adalah matriks $\boldsymbol{X}$ dan vektor $\boldsymbol{y}$.
b. Untuk a = 1,...,h dengan h adalah banyak skor komponen PLS. Pada $a = 1$, tetapkan $\boldsymbol{X}_0 = \boldsymbol{X}$ dan $\boldsymbol{y}_0 = \boldsymbol{y}$ dan lakukan langkah-langkah berikut:
   1) Berdasarkan persamaan (2.9) maka pembobot $\boldsymbol{w}_{\mathbf{a}}$ dihitung menggunakan persamaan:

$$w_a = \frac{X_{a-1}^T \cdot y_{a-1}}{\|X_{a-1}^T \cdot y_{a-1}\|} \tag{2.11}$$

   di mana penyebut pada persamaan (2.11) menunjukkan normalisasi dari matriks $(X_{a-1}^T \cdot y_{a-1})$ yaitu akar dari jumlah kuadrat bagi elemen-elemen matriks $(X_{a-1}^T \cdot y_{a-1})$.
   2) Menghitung skor komponen t_a

$$t_a = X_{a-1}w_a$$

   3) Menghitung faktor *loading* dari **X** yaitu:

$$p_a = \frac{X_{a-1}^T \cdot t_a}{t_a^T\, t_a}$$

   4) Menghitung faktor *loading* dari **y** yaitu:

$$q_a = \frac{y_{a-1}^T \cdot t_a}{t_a^T\, t_a}$$

5) Menghitung residual dari model matematis regresi PLS, yang selanjutnya nilai residual tersebut digunakan sebagai matriks masukan **X** dan vektor **y** yang baru

$$E_a = X_{a-1} - t_a p_a^T = X_a$$

$$f_a = y_{a-1} - q_a\, t_a = y_a$$

dan dilakukan iterasi selanjutnya untuk $a \leftarrow a + 1$

c. Didapatkan model regresi laten dari regresi PLS yaitu

$$\hat{y} = T_h\, q_h \tag{2.12}$$

Vektor l*oading* q_h yang berukuran h×*1* merupakan koefisien regresi yang didapatkan dari hasil meregresikan variabel respon **y** terhadap skor komponen T_h menggunakan penduga OLS. Jika model persamaan regresi laten (2.12) dikembalikan ke persamaan umum regresi dengan koefisien regresi yang digunakan berukuran p×*1* dan dinotasikan dengan $\hat{\beta}_h^{pls}$ maka $\hat{\beta}_h^{pls}$ dapat dicari melalui:

$$\hat{\beta}_h^{pls} = W_h q_h \tag{2.13}$$

Untuk $W_h = (r_1, \ldots, r_h)$, $P_h = (p_1, \ldots, p_h)$

Selanjutnya berdasarkan persamaan (2.12) dan (2.13) maka didapatkan model akhir regresi PLS adalah:

$$\hat{y}_h = X\hat{\beta}_h^{pls}$$

2.5 *Mean Squared Error* (MSE)

Dalam statistik, *Mean Squared Error* (MSE) sebuah estimator adalah nilai yang diharapkan dari kuadrat *error*. *Error* yang ada menunjukkan seberapa besar perbedaan hasil estimasi dengan nilai yang akan diestimasi. Perbedaan itu terjadi karena adanya keacakan pada data atau karena estimator tidak mengandung informasi yang dapat menghasilkan estimasi yang lebih akurat.

Terdapat beberapa metode yang biasa digunakan untuk menguji model manakah yang lebih baik. Namun terkadang dua metode yang berbeda dapat memberikan hasil yang berbeda, sehingga tidak ada ketetapan khusus metode manakah yang lebih baik dalam mengukur kecocokan model. Peneliti menggunakan metode MSE untuk menentukan model manakah yang lebih baik. Nilai MSE diperoleh dengan membagi *Sum Square Error* (SSE) dengan jumlah data atau dalam bentuk matematis dapat ditulis sebagai berikut:

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

dimana y merupakan variabel respon, $\hat{y}$ merupakan prediktor dan *n* adalah banyaknya data. Model dengan nilai MSE yang mendekati angka nol mengindikasikan bahwa model tersebut merupakan model yang lebih baik.

2.6 Tranformasi Komponen *Partial Least Square* (PLS) ke Variabel Asli

Dari persamaan (2.10) selanjutnya dapat dituliskan ke dalam bentuk variabel aslinya, yaitu:

$$\begin{aligned} Y &= \sum_{h=1}^{m} q_h t_h + \varepsilon \\ &= \sum_{h=1}^{m} q_h \left(\sum_{j=1}^{p} W_{(h)j} X_j\right) + \varepsilon \\ &= \sum_{j=1}^{p} \sum_{h=1}^{m} q_h W_{(h)j} X_j + \varepsilon \\ Y &= \sum_{j=1}^{p} b_j X_j + \varepsilon \end{aligned}$$

Dimana:

Y	= variabel respon
q_h	= koefisien regresi Y terhadap t_h
$t_h = \sum_{j=1}^{p} W_{(h)j} X_j$	= komponen utama ke-h yang tidak saling berkorelasi, $(h = 1,2, \dots, m)$
$b_j = \sum_{h=1}^{m} q_h W_{(h)j}$	= vektor koefisien regresi Y terhadap variabel X_j
$W_{(h)j}$	= koefisien bobot untuk variabel X_j pada komponen utama PL ke- h
ε	= vektor eror

BAB 3 METODOLOGI PENELITIAN

3.1 Sumber Data

Dalam penelitian ini data yang digunakan adalah data primer yang diperoleh dari data pengecekan kesehatan secara berskala siswa-siswi TK Pertiwi Jember tahun 2014–2015. Data yang digunakan untuk penelitian PLS berupa data tinggi badan, berat badan, lingkar kepala, lingkar lengan, dan umur siswa.

3.2 Variabel Penelitian

Variabel penelitian meliputi variabel respon (Y) dan variabel penjelas (X), dimana variabel respon (Y): tinggi badan (TB) dan variabel penjelas (X) : X_1 = lingkar kepala (LK), X_2 = lingkar lengan (LL), X_3 = umur (UMUR), X_4 = berat badan (BB).

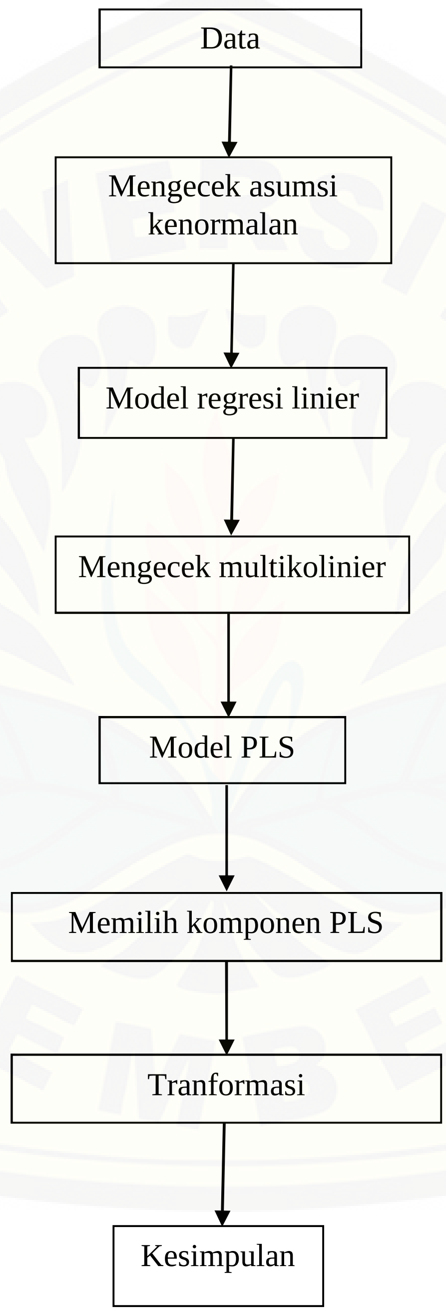
3.3 Pengolahan Data

Langkah-langkah yang akan dilakukan dalam penyelesaian skripsi ini adalah sebagai berikut:

a. Pada awal didefinisikan terlebih dahulu data tinggi badan dan data yang mempengaruhi tinggi badan.
b. Mengecek asumsi kenormalan dengan menggunakan metode *Shapiro-Wilk* melalui program *R* dengan scipt > shapiro.test(tk_pertiwi3$TB)
c. Membentuk persamaan model regresi linier seperti persamaan (2.1).
d. Mengecek nilai VIF dengan rumus VIF seperti persamaan (2.6).
e. Membentuk model *Partial Least Square* (PLS) dengan menggunakan program *R* dengan script > reg.pls$coefficients.
f. Memilih hasil komponen dari model *Partial Least Square* (PLS).

g. Mentransformasi dari model *Partial Least Square* (PLS) ke model regresi linier.

Berikut adalah skema proses penelitian dari metode analisis pada Gambar 3.1



Gambar 3.1 Skema metode penelitian

3.4 Pengolahan data dengan menggunakan program R

Untuk pengolahan data digunakan bantuan sofware R version 3.1.0 dengan paket `pls` . Program R merupakan program opensource yang dapat diperoleh melalui situs http://www.r-project.org.

Paket `pls` adalah paket yang digunakan untuk melakukan regresi parsial kuadrat terkecil (Plsr) atau komponen regresi utama (PCR), dengan rumus. Paket `pls` dapat digunakan untuk mencari atau menghitung Cross-validasi, Metode prediksi, ekstraksi model, plot, cetak dan ringkasan. Fungsi yang dipakai pada paket pls adalah plsr.

Untuk Fungsi R nya adalah sebagai berikut:

```
mvr(formula, ncomp, data, subset, na.action,
method = pls.options()$mvralg,
scale = FALSE, validation = c("none", "CV", "LOO"),
model = TRUE, x = FALSE, y = FALSE, ...)
plsr(..., method = pls.options()$plsralg)
pcr(..., method = pls.options()$pcralg)
```

fungsi diatas dijelasakan sebagai berikut:

`formula` sebuah model atau formula yang mendukung konstruksi lm

`ncomp` jumlah komponen yang akan disertakan dalam model

`data` frame data opsional untuk menentukan data agar sesuai dengan `model`

`subset` vektor opsional untuk menentukan subset dari pengamatan yang akan digunakan dalam proses fitting.

`na.action` fungsi yang menunjukkan apa yang harus dilakukan ketika data mengandung nilai-nilai yang hilang. Standarnya diatur oleh pengaturan na.action , dan na.fail jika itu tidak diset. Nilai lain yang