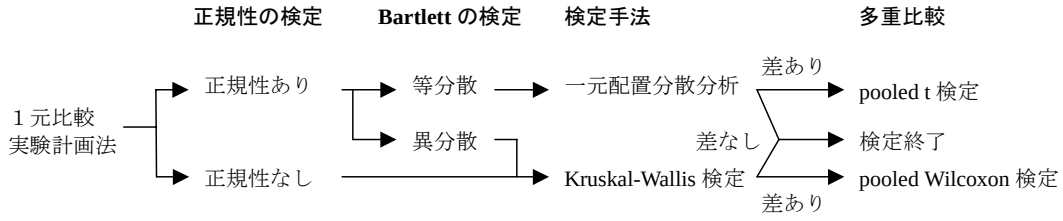
演習１　実験計画法



ある４つの中学について英語・数学・国語の試験結果を調べた。多変量演習 1.txt のデータを読み込んで、以下の問題に答えよ。但し、検定は有意水準 5%とすること。

１．中学

１）A 中学　２）B 中学　３）C 中学　４）D 中学

２．英語点数

３．数学点数

４．国語点数

問題

１）各中学、各教科の平均値を求めよ。

	A 中学	B 中学	C 中学	D 中学	全体
英語					
数学					
国語					

２）各中学、各教科の中央値を求めよ。

	A 中学	B 中学	C 中学	D 中学	全体
英語					
数学					
国語					

３）数学について、すべての中学の分布は正規分布といえるか。

正規分布と［みなす・いえない］。

４）数学について、各中学の分散に差があるといえるか。（調べられる場合のみ）

検定確率［　　］　［等分散・異分散］とみなす。

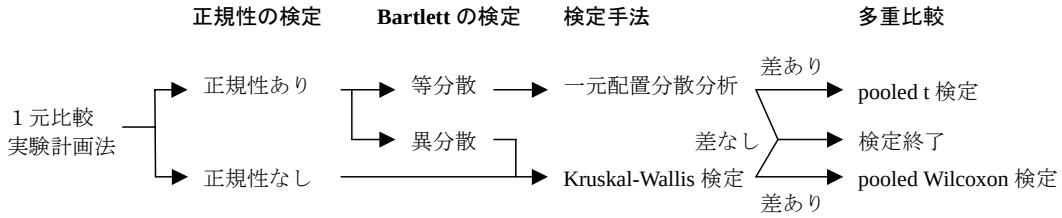
５）数学について、各中学の平均（中央）値に差があるといえるか。

検定名［　　］　検定確率［　　］

判定　平均（中央）値に差があると［いえる・いえない］。

6）数学について各中学の平均（中央）値に差があるとすると、どの中学の間に差があるか調べよ。(調べられる場合のみ)
検定名［　　］結果［　　］
7）数学の成績の良い順に中学を並べよ。但し、統計的に差があるものは不等号で、統計的に差がないものは等号で表し、同順位とみなすこと。(調べられる場合のみ)
［　　］
8）国語について、すべての中学の分布は正規分布といえるか。
正規分布と［みなす・いえない］。
9）国語について、各中学の分散に差があるといえるか。(調べられる場合のみ)
検定確率［　　］　［等分散・異分散］とみなす。
10）国語について、各中学間の平均（中央）値に差があるといえるか。
検定名［　　］　検定確率［　　］
判定　平均（中央）値に差があると［いえる・いえない］。
11）国語について、各中学間の平均（中央）値に差があるとすると、どの中学の間に差があるか調べよ。(調べられる場合のみ)
検定名［　　］結果［　　］
12）国語の成績の良い順に中学を並べよ。但し、表示法は数学の場合に習え。
［　　］
13）３教科の分布はすべて正規分布といえるか。
正規分布と［みなす・いえない］。
14）正規分布の場合、３教科の分散は差があるといえるか。
検定確率［　　］　［等分散・異分散］とみなす。
15）３教科の平均（中央）値に差があるといえるか。対応は考えないものとせよ。
検定名［　　］　検定確率［　　］
判定　平均（中央）値に差があると［いえる・いえない］。
16）３教科の平均（中央）値に差があるとすれば、どの教科の間に差があるか。
検定名［　　］結果［　　］
17）点数の良い順に教科を並べよ。但し、表示法は数学の場合に習え。
［　　］

演習２　実験計画法２



ある商品の売り上げ（万円）を４つの地域で規模の同じコンビニを対象に調査した。これらの売り上げに地域差はあるといえるか、またあるとするとどの地域間にあるか。多変量演習 2.txt のデータを読み込み、以下の問題に答えよ。また結果は有意水準 5%で判定せよ。

問題

１）各都市の売り上げの平均値と中央値を求めよ。

	東京	名古屋	大阪	福岡
平均値				
中央値				

２）各都市の売り上げの分布はすべて正規分布といえるか。

正規分布と［みなす・いえない］。

３）正規分布の場合、売り上げの分散は等しいといえるか。

検定名［　　］　検定確率［　　］

判定　［等分散・異分散］とみなす。

４）各都市の売り上げの平均（中央）値間に差があるといえるか。

検定名［　　］　検定確率［　　］

判定　平均（中央）値に差があると［いえる・いえない］。

５）売り上げの平均（中央）値に差があるとすれば、どの都市の間に差があるか。多重比較の検定確率を表示せよ。

検定名［　　］

	東京	名古屋	大阪	福岡
東京	1			
名古屋		1		
大阪			1	
福岡				1

６）上の結果から差のある都市名をｘｘ＜ｘｘというように平均値の大小の不等号ですべて示せ。［　　］

以下は正しいと思われる検定４）、５）と結果を比較するための計算である。

７）各都市の売り上げの中央（平均）値間に差があるといえるか。４）と異なる検定を用いて判定せよ。

検定名［　　］　検定確率［　　］

判定　中央（平均）値に差があると［いえる・いえない］。

８）売り上げの中央（平均）値に差があるとすれば、どの都市間に差があるか。５）と異なる検定を用いて多重比較の検定確率を表示せよ。

検定名［　　］

	東京	名古屋	大阪	福岡
東京	1			
名古屋		1		
大阪			1	
福岡				1

９）都市の売り上げの平均（中央）値に差があるかどうか、ｔ検定を用いて各２群間の差の検定確率を求め、以下の表に記入せよ。

	東京	名古屋	大阪	福岡
東京	1			
名古屋		1		
大阪			1	
福岡				1

10）各都市の売り上げの中央（平均）値に差があるかどうか、Wilcoxon の順位和検定を用いて各２群間の検定確率を求め、以下の表に記入せよ。

	東京	名古屋	大阪	福岡
東京	1			
名古屋		1		
大阪			1	
福岡				1

11）正規性・等分散性が認められる場合、一元配置分散分析と Kruskal-Wallis 検定ではどちらが差を見出し易いか。［一元配置分散分析・Kruskal-Wallis 検定］

12）pooled 統計量を用いた多重比較と通常の検定とではどちらが差を見出し易いか。［pooled 統計量・通常の検定・どちらともいえない］

演習３　重回帰分析１

解説

データ Samples¥重回帰分析 1.txt を用いて、体重を身長と胸囲の 1 次関数で予測する。

体重 ＝ b_1 身長＋b_2 胸囲＋b_0　　の形で体重を予測する。

目的変数：体重　　説明変数：身長，胸囲

係数の値は？　　→　偏回帰係数

説明変数の重要性は？　　→　標準化偏回帰係数

どの程度予測できるか？　　→　重相関係数，寄与率

このモデルは有効か？　　→　F 検定値と確率（要残差正規性）

それぞれの係数は有効か？　　→　t 検定値と確率（要残差正規性）

他の変数の影響を除いた目的変数と各説明変数の相関は？　　→　偏相関係数

どの程度予測できているのか図的に見たい　　→　散布図

どの程度予測できているのかデータ毎に見たい　　→　予測値と残差

解答例

目的変数を体重に、説明変数を身長と胸囲にして、重回帰分析を行ったところ、以下の回帰式を得た。

体重 = 0.3861*身長+0.8575*胸囲-80.7427

予測体重と実測体重の相関である重相関係数は 0.84055 で、回帰式の寄与率は 0.70652 となった。これから体重変動の約 71%が説明できることが分かる。この実測体重と予測体重の関係を散布図にすると、縦軸を実測体重として、以下のように表される。

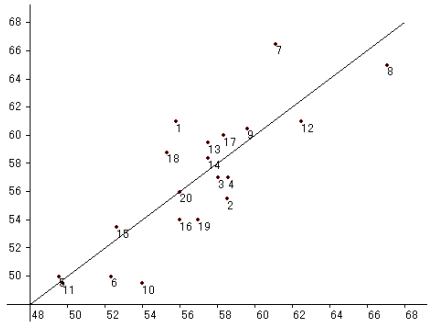


図　実測値（縦軸）／予測値（横軸）の散布図

また回帰式の妥当性の検定を行ったところ p=0.00003 となり、妥当性が有意に示された。

各変数の予測における重要性を示す標準化偏回帰係数は、身長が 0.4333、胸囲が

0.6401 と胸囲が少し上回っている。

各偏回帰係数が 0 と異なることを示す検定では､身長が p=0.00488､胸囲が p=0.00018､切片は p=0.00233 となり、各係数とも有意に 0 と異なっている。

以上のことからこの回帰式は予測モデルとして、かなり良いモデルになっている。

問題１

多変量演習 3.txt のデータについて、試験成績を目的変数に他の変数を説明変数にして重回帰分析を行い、結果を上の例にならってまとめよ。

問題２

１）問題１の重回帰分析で、全変数を使った重回帰式はどのように与えられるか。

試験成績＝［　　］×評定平均＋［　　］×模試１

［　　］×模試２＋［　　］×模試３＋［　　］

２）この重回帰式の寄与率はいくらか。［　　］

偏回帰係数の有効性（0 でない）の検定で、確率の大きい順に変数を除いて行き、すべての偏回帰係数が有効である回帰モデルを作り、以下の問いに答えよ。

３）この回帰モデルについて以下の値を求めよ。但し変数名はこのモデルで残ったものだけで残りは空欄にしておくこと。

変数名				
偏回帰係数				
標準化偏回帰係数				
検定確率値				

４）重回帰式はどのようになったか。

試験成績＝［　　］

５）寄与率はいくらになったか。［　　］

６）全変数の場合の寄与率に比べてかなり低くなったと思うか。[思う・思わない]

７）上の重回帰式を新しい予測モデルにして良いと思うか。[思う・思わない]

８）データの中で最初の学生の予測試験成績はいくらか。［　　］

９）その学生の実際の試験成績と予測試験成績との差はいくらか。［　　］

１０）　上の重回帰式を利用すると以下の点数を取った学生の試験成績は何点に予測されるか。

変数名	評定平均	模試１	模試２	模試３
成績	3.5	70	73	75

予測試験成績［　　］

演習４　重回帰分析２

解説

目的変数＝b_1 説明変数１＋b_2 説明変数２＋・・・＋b_0　　の形で予測する。

係数の値は？　　→　偏回帰係数

説明変数の重要性は？　　→　標準化偏回帰係数

どの程度予測できるか？　　→　重相関係数，寄与率

このモデルは有効か？　　→　Ｆ検定値と確率（要残差正規性）

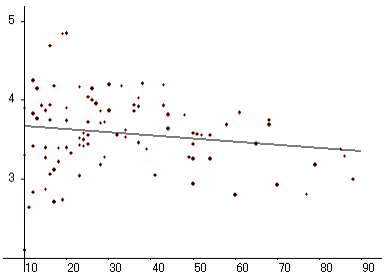
それぞれの係数は有効か？　　→　ｔ検定値と確率（要残差正規性）

どの程度予測できているのか図的に見たい　　→　散布図

どの程度予測できているのかデータ毎に見たい　　→　予測値と残差

問題

多変量演習 4.txt のデータは各質問項目について 5 段階評価で、講義ごとに平均を取ったものである。基本統計の相関と回帰分析及び、多変量解析の重回帰分析を用いて以下の問いに答えよ。



総合評価を調査数で予測する回帰モデル

１）総合評価を縦軸、調査数を横軸とした右上の散布図を描け。

２）総合評価と調査数の相関係数を求めよ。［　　］

３）回帰式（直線の方程式）を求めよ。

総合評価＝［　　］調査数＋［　　］

４）この式から調査数が 50 人増えると総合評価はいくら減るか。［　　］

５）回帰式の寄与率を求めよ。［　　］

６）この回帰式は予測モデルとして有効か。[有効である・有効でない]

総合評価を調査数以外のすべての変数で予測する重回帰モデル

７）回帰式を求めよ。

総合評価＝　［　　］進む速さ　＋［　　］声の大きさ

＋［　　］黒板等　　＋［　　］私語注意

＋［　　］分かり易さ＋［　　］有益さ

＋［　　］受講態度　＋［　　］

8）この回帰式の寄与率を求めよ。［　　］

9）回帰式の係数の t 検定（偏回帰係数が 0 と異なるかどうかの検定）の確率値が 0.05 を超えるものの中で最大となる変数（最も不要な変数）を順次削除していくと、最終的に残るものは何か。各段階の検定確率値を記入せよ。但し、削除した変数のところは以後空欄にし、すべての確率が 0.05 未満になった場合は確定とする。

	7 変数	6 変数	5 変数	4 変数
進む速さ				
声の大きさ				
黒板等				
私語注意				
分かり易さ				
有益さ				
受講態度				

10）最終的な回帰式はどのようになるか。不要な変数の係数欄は空欄のままでよい。

総合評価＝　［　　］進む速さ　＋［　　］声の大きさ
＋［　　］黒板等　　＋［　　］私語注意
＋［　　］分かり易さ＋［　　］有益さ
＋［　　］受講態度　＋［　　］

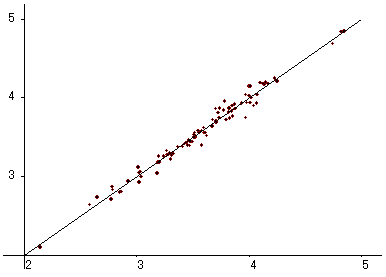
11）上のような処理は正しいと思われるか。［正しい・少し注意が必要］

12）上の回帰式の寄与率を求めよ。［　　］

13）上の回帰式の寄与率はすべての変数を使った場合に比べ大きく下がっているか。
［大きく下がっている・あまり下がっていない］

14）この式を新しい予測モデルとして採用するか。
［採用する・採用しない］

15）予測値がどの程度実測値に近いかを見るために、右のような散布図を描け。



16）総合評価に影響を与える重要な説明変数を２つ挙げよ。［　　］［　　］

17）データ中の最初（1 番）の授業について、総合評価の実測値，その予測値，残差（実測値と予測値の差）はいくらか。

実測値［　　］　予測値［　　］　残差［　　］

18）すべての質問項目の値が 3.5 の授業の総合評価はいくらに予測されるか。
［　　］

演習５　判別分析１

データ Samples¥判別分析 1.txt を用いて、試験の合否を勉強時間とそれまでの試験の平均点の１次関数で判別する。

判別分析の目的

2 群（多群）を判別する最適な 1 次式を求める。

判別値＝b_1 勉強時間＋b_2 平均点＋b_0　　判別関数

判別分析が有効に利用できる条件は？　→　正規性，等共分散性（等共分散の検定）

判別関数の係数は？　　→　判別関数の欄

判別関数で群を分けるのは？　→　判別の分点 0（多群の場合は値が最大の群）

各係数の有効性は？　→　確率の欄（係数が 0 と異なるかの検定）

誤判別の程度は？　　→　誤判別確率（実測と理論）

マハラノビス距離とは　　→　どの程度 2 群が離れているかを表わす指標

データ毎の判別関数の値と判別状況　　→　判別得点

事象の生起確率とは？　→　合格・不合格の現れる確率が大きく異なっている場合の措置，各群同じかデータ数からが実用的

誤判別損失とは？　→　間違った判断をした場合の致命傷の程度
大きな差がない限り、各群 1 とするが実用的

解答例

正規性の検定から、2 群とも正規性があるとみなされ、等共分散性の検定でも共分散に差があるとは言えなかった。以上から判別分析が適用可能であると判断した。

2 群の生起確率を同じとし、誤判別損失を等しいとすると、判別分析によって、以下の判別関数が得られた。

y=2.2461*勉強時間+0.2007*平均点-23.0187

データはこの判別関数の値をもとに、判別の分点を 0 として、2 群に分けられる。

各係数については、勉強時間が p=0.00013、平均点が p=0.00061 のように、両方とも有意に 0 でないことが示された。このことから２つの変数とも有効であると思われる。

マハラノビス距離 5.6823 から、理論的な誤判別確率として p=0.117 が予想される。また、実際に判定を行うと、1 群を 2 群と間違える割合が 7.7%、その逆が 5.9%となる。これらの数値から、判別はかなりうまく行われたものと思われる。

問題１

多変量演習 5.txt のデータについて、合否を他の変数で予測する判別分析を行い、結果を上の例にならってまとめよ。合否の欄で、1 は合格、2 は不合格である。

問題２

問題１のデータを用いて、生起確率をデータ数から、誤判別損失を各群１として判別分析を行い、以下の問いに答えよ。

１）このデータに判別分析は有効に利用可能か？

正規性の検定　　正規性があると［みなす・いえない］

等共分散性　　検定確率［　　］，等共分散と［みなす・いえない］

判別分析は［利用可能・要注意］

２）判別関数を求めよ。

判別値＝　［　　］内申＋［　　］模試１

＋［　　］模試２＋［　　］

３）判別の分点［　　］

４）実測値から求めた誤判別の確率は？

合格を不合格と［　　］　不合格を合格と［　　］

５）上の誤判別でどちらの場合が損失が大きいと思われるか。

［合格を不合格・不合格を合格］と誤判別する場合

６）これに従って、誤判別損失の値を合格を不合格と判定したとき 1，不合格を合格と判定したとき 2 としたい。そのときの実測値から見た誤判別の確率はどうなるか。

合格を不合格と［　　］　不合格を合格と［　　］

７）元の設定で、各係数の有効性の検定で、5%の有意水準で有意でない変数はどれか。

変数［　　］　検定確率［　　］

８）その変数を取り除いて再度判別分析を行い、判別関数を求めよ。但し、取り除いた変数のところは空欄とせよ。

判別値＝　［　　］内申＋［　　］模試１

＋［　　］模試２＋［　　］

９）この場合、実測値から見た誤判別の確率はどうなるか。

合格を不合格と［　　］　不合格を合格と［　　］

10）元のモデルとこの新しいモデルとで誤判別確率に大きな差があると思われるか。

［大きな差がある・大した差ではない］と思われる。

11）新しいモデルで、先頭（1 番）の人の判別値はいくらか。［　　］

12）新しいモデルで、内申 3.4 点，模試１ 65 点，模試２ 70 点の人の判別値はいくらか、またその人の合否を判定せよ。

判別値［　　］　判定［合格・不合格］

演習６　判別分析２

データ Samples¥判別分析 1.txt を用いて、試験の合否を勉強時間とそれまでの試験の平均点の１次関数で判別する。

判別分析の目的　　２群（多群）を判別する最適な１次式を求める。

２群の場合　　判別得点＝b_1 勉強時間＋b_2 平均点＋b_0　　判別関数

判別の分点 0 より大きいか小さいかで１群と２群を分ける

２群以上の場合　判別得点＝b_1 勉強時間＋b_2 平均点＋b_0－判別の分点

判別得点が最大となる群に属すると判定する。

判別分析が利用できる条件は？　→　正規性，等共分散性（等共分散の検定）

判別関数の係数は？　　→　判別関数の欄

判別関数で群を分けるのは？　→　判別の分点（多群の場合は値が最大の群）

各係数の有効性は？　→　確率の欄（係数が 0 と異なるかの検定）

誤判別の程度は？　　→　誤判別確率（実測と理論）

マハラノビス距離とは　　→　どの程度 2 群が離れているかを表わす指標

データ毎の判別関数の値と判別状況　　→　判別得点

事象の生起確率とは？　→　合格・不合格の現れる確率が大きく異なっている場合の措置，各群同じかデータ数からが実用的

誤判別損失とは？　→　間違った判断をした場合の致命傷の程度

大きな差がない限り、各群１とするが実用的

多変量演習 6.txt のデータはある職業の適性について調べた結果である。適性は、１．適性あり、２．努力しだい、３．適性なしに分類され、それを予測するデータとして回答者の年齢、学力テスト、体力テスト、面接（10 段階）の結果が含まれている。

１ページ目はすべてのデータで、２ページ目は努力しだいを取り除いたものである。

問題１

２ページ目のデータを用いて、生起確率をデータ数から、誤判別損失を各群１として判別分析を行い、以下の問いに答えよ。

１）このデータに判別分析は利用可能か？

正規性の検定　　正規性があると［みなす・いえない］

等共分散性　　検定確率［　　］，等共分散と［みなす・いえない］

判別分析は［利用可能・要注意］

２）判別関数を求めよ。

判別得点＝［　　］年齢＋［　　］学力テスト
＋［　　］体力テスト＋［　　］面接＋［　　］

3）判別の分点［　　］

4）実測値から求めた誤判別確率は。
適性ありをなしと［　　］　適性なしをありと［　　］

5）年齢 30 歳、学力テスト 55 点、体力テスト 45 点、面接 6 点の人はどちらに判定されるか。判別関数値［　　］　判別［適性あり・適性なし］

6）この適性判定が、有能な人間はぜひ採用したいという入社試験で使われた場合、会社にとってどちらの誤判別損失が大きいと思われるか。
［適性ありをなし・適性なしをあり］と誤判別する場合

7）誤判別損失の値の小さい方を 1、大きい方を 2 とした場合、実測値から見た誤判別の確率はどうなるか。
適性ありをなしと［　　］　適性なしをありと［　　］

8）この結果は誤判別損失が等しいとした場合と比べて、6）の会社にとって改善されたか。［改善された・改善されていない］

問題２

１ページ目のデータを用いて、生起確率をデータ数から、誤判別損失を各群１として判別分析を行い、以下の問いに答えよ。

1）３つの判別得点の式を求めよ。但し定数項は判別の分点を引いたものとする。
判別得点１＝［　　］年齢＋［　　］学力テスト
＋［　　］体力テスト＋［　　］面接＋［　　］
判別得点２＝［　　］年齢＋［　　］学力テスト
＋［　　］体力テスト＋［　　］面接＋［　　］
判別得点３＝［　　］年齢＋［　　］学力テスト
＋［　　］体力テスト＋［　　］面接＋［　　］

2）実測値から求めた誤判別確率はいくらか。　適性ありを他と［　　］
努力しだいを他と［　　］適性なしを他と［　　］

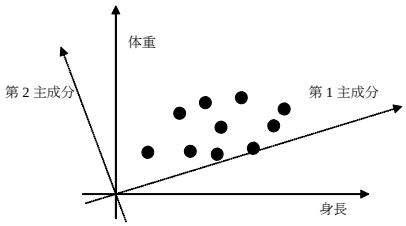
3）先頭の人の３つの判別得点を求めよ。
判別得点１［　　］判別得点２［　　］判別得点３［　　］

4）年齢 35 歳、学力テスト 50 点、体力テスト 50 点、面接 6 点の人はどれに判定されるか。［適性あり・努力しだい・適性なし］

5）２ページ目のデータを用いて、２群の判別関数と多群の判別関数の関係を考えよ。

演習７　主成分分析１

Samples¥主成分分析 1.txt のデータから、変数の１次関数として体格を表す特徴的な指標を作る。



主成分分析の目的

複数の変数を１次関数として組み合わせて、いくつかの特徴的な量を作り出す。

各主成分の係数値は？　→　固有ベクトルの値（全体的に符号を変えてもよい）

各主成分のばらつき（分散）は？　→　各主成分の固有値

各主成分の重要性（分散の割合）は？　→　各主成分の寄与率

各主成分と各変数の関係は？　→　因子負荷量（各主成分と各変数の相関係数）

何番目の主成分まで意味があるか？　→　等固有値の検定（要正規性）

主成分が意味がある→他の主成分と値が異なる

データごとの主成分の値は？　→　主成分得点

共分散行列からと相関行列からどちらを使う　→　実用的には相関行列が一般的

まとめ

変数に身長、体重、胸囲、座高の４つをとって主成分分析を行なった。各変数の値に大きな差がないことから、ここでは共分散行列を基にした方法を用いている。変数は正規分布するものとみなされ、等固有値の検定も利用可能である。

第１主成分は１次式の係数の値（固有ベクトルの値）がすべて正であることから身体の大きさを表わす変数であると考える。また、第 2 主成分は身長・座高と体重・胸囲で符号が違うことから、肥満の程度を表わす変数であると考える。

これらの主成分の寄与率をみると、第１主成分が 0.8914 と非常に大きく、他はすべて 0.08 以下になっている。また等固有値の検定より、第１主成分と第２主成分が利用可能であることが分かる。それ以降の主成分については意味付けも困難であり、利用しない。最後に結果を式で表わしておく。

身体の大きさを表わす主成分

第１主成分＝0.6240 身長＋0.5592 体重＋0.4083 胸囲＋0.3622 座高

肥満の程度を表わす主成分

第２主成分＝－0.6456 身長＋0.3456 体重＋0.6605 胸囲－0.1660 座高

問題１

多変量演習 7.txt のデータについて、特徴的な量を変数の１次式で表す主成分分析を行い、結果を上の例にならってまとめよ。

問題２

多変量演習 7.txt のデータについて、共分散行列をもとにするモデルを用いて以下の問いに答えよ。

１）各主成分の固有値（分散の値)、寄与率、累積寄与率を求めよ。

	第１主成分	第２主成分	第３主成分	第４主成分
固有値				
寄与率				
累積寄与率				

２）各変数の正規性の検定　正規分布と［みなす・いえない］

これより等固有値の検定は［利用可能・利用不可能］

３）等固有値の検定が利用できる場合、有意に固有値が異なるといえる主成分の数は

［　　］個

４）これらの主成分で説明できるのは全体の変動の何%か。［　　］%

５）第１主成分と第２主成分の関数はどのように表されるか。

第１主成分＝　［　　］国語＋［　　］算数

＋［　　］理科＋［　　］社会

第２主成分＝　［　　］国語＋［　　］算数

＋［　　］理科＋［　　］社会

６）２つの主成分と各変数との相関係数を求めよ。

相関係数	国語	算数	理科	社会
第１主成分				
第２主成分				

７）これら２つの主成分はどのように意味づけられるか。

第１主成分　意味［　　］

第２主成分　意味［　　］

８）先頭（1 番）の生徒の軸の平行移動をした２つの主成分得点を求めよ。

第１主成分得点［　　］　第２主成分得点［　　］

９）２つの主成分の意味を考えて、この生徒にはどんな特徴があるか。

［　　］

10）主成分得点で軸の平行移動を行わない場合と行った場合の違いは。

行った主成分得点＝行わない主成分得点－［　　］

演習８　主成分分析２

主成分分析の目的

　複数の変数を１次関数として組み合わせて、いくつかの特徴的な量を作り出す。

各主成分の係数値は？　→　固有ベクトルの値（全体的に符号を変えてもよい）

各主成分のばらつき（分散）は？　→　各主成分の固有値

各主成分の重要性（分散の割合）は？　→　各主成分の寄与率

各主成分と各変数の関係は？　→　因子負荷量（各主成分と各変数の相関係数）

何番目の主成分まで意味があるか？　→　等固有値の検定（要正規性）

主成分が意味がある→他の主成分と値が異なる

データごとの主成分の値は？　→　主成分得点

共分散行列からと相関行列からどちらを使う　→　実用的には相関行列が一般的

問題１

　多変量演習 8.txt のデータはある学校で測定した小学６年生の運動適性テストの結果である。相関行列を用いたモデルで主成分分析を行い、以下の問いに答えよ

１）変数間の共分散行列を求めよ。但し、数値は標準的な形に直して表せ。

	立幅跳び	腹筋	腕立伏せ	往復走	５分間走
立幅跳び					
腹筋					
腕立伏せ					
往復走					
５分間走					

２）変数間の相関行列を求めよ。

	立幅跳び	腹筋	腕立伏せ	往復走	５分間走
立幅跳び					
腹筋					
腕立伏せ					
往復走					
５分間走					

３）どの種目間の相関が最も高いか。［　　］と［　　］

４）各変数の平均値と標準偏差（不偏分散からのもの）を求めよ。

	立幅跳び	腹筋	腕立伏せ	往復走	５分間走
平均値					
標準偏差					

5）相関行列は、各変数を以下の式のように標準化した共分散行列に等しい。以下の値のデータを各変数ごとに標準化せよ。標準化した値＝（値－平均値）／標準偏差

	立幅跳び	腹筋	腕立伏せ	往復走	５分間走
値	190	30	30	40	1120
標準化値					

6）上の値の人はどの種目が最も優れているか。［　　］

7）各主成分の固有値（分散の値)、寄与率、累積寄与率を求めよ。

	第１主成分	第２主成分	第３主成分	第４主成分	第５主成分
固有値					
寄与率					
累積寄与率					

8）各変数の正規性の検定　正規分布と［みなす・いえない］
これより等固有値の検定は［利用可能・利用不可能］

9）等固有値の検定が利用できる場合、有意に固有値が異なるといえる主成分の数は［　　］個（これは目安と考える）

10）上から２つの主成分で説明できるのは全体の変動の何%か。［　　］%

11）２つの主成分関数はどのように表されるか。（但し相関行列のモデルの場合、各変数は標準化したものを用いること）
第１主成分＝［　　］立幅跳び＋［　　］腹筋
＋［　　］腕立伏せ＋［　　］往復走＋［　　］５分間走
第２主成分＝［　　］立幅跳び＋［　　］腹筋
＋［　　］腕立伏せ＋［　　］往復走＋［　　］５分間走

12）これらの主成分はどのように意味づけられるか。
第１主成分　意味［　　］
第２主成分　意味［　　］

13）相関行列のモデルでは標準化したデータを用いることに注意して、５）で与えた生徒の２つの主成分得点を求めよ。
第１主成分得点［　　］　第２主成分得点［　　］

14）２つの主成分の意味を考えて、この生徒にはどんな特徴があるか。
［　　］

15）相関行列から始めた場合、軸の平行移動（主成分得点の平均を引く操作）を行う場合と行わない場合で差があるか。　　［ある・ない］

16）15）の理由はどうしてと考えられるか。
標準化したデータ及び主成分得点の［　　］が［　　］となるから。

演習９　因子分析

解説

因子分析の目的　各変数の背後にある共通因子を求め、それらの１次関数として各変数が表されるように係数を求める。

各因子の係数値は？　→　因子負荷量の値（全体的に符号を変えて見てもよい）

各因子と各変数の相関係数は？　→　因子負荷量の値（因子間は無相関とした場合）

各因子の重要性は？　→　各因子の寄与率

何番目の因子まで考えるか？　→　累積寄与率が 90%程度まで

相関行列の固有値で１より大きい固有値の数

データごとの因子の値は？　→　因子得点

因子の値を求めるときの係数の値は？（変数は標準化）　→　因子得点係数の値

問題

多変量演習 7.txt はある小学校における４教科の試験の成績である。因子分析を用いて特徴を分析し、以下の問いに答えよ。

１）各科目間の相関行列の固有値を大きい順に求めよ。

1	2	3	4

２）因子数を３として、因子分析を行い、寄与率を求めよ。

因子１	因子２	因子３

３）これらのデータから因子数はいくつと決めるのが妥当か。［　　］個

以後因子数を２つと決めて各質問に答えよ。

４）各因子の因子負荷量を求めよ。

回転なしの場合

	国語	算数	理科	社会
第１因子				
第２因子				

回転ありの場合

	国語	算数	理科	社会
第１因子				
第２因子				

５）この場合の各因子の意味を解釈せよ。

回転なしの場合

第１因子：[　　]を表す因子

第２因子：[　　]を表す因子

回転ありの場合

第１因子：[　　]を表す因子

第２因子：[　　]を表す因子

以後はバリマックス回転ありとして質問に答えよ。

６）先頭から３人の因子の値（因子得点）を推定せよ。

	第１因子	第２因子
1		
2		
3		

７）１番の人にはどんな特徴があるか。

[　　]

８）因子得点を求める際の係数を求めよ。但し、変数は標準化されているものとする。

	国語	算数	理科	社会
第１因子				
第２因子				

９）国語について最初の３人の標準化された実測値と因子得点から求められる予測値を求めよ。

	実測値	予測値
1		
2		
3		

10）各教科の実測値と予測値の相関係数を求めよ。

国語	算数	理科	社会

11）予測値が２つの因子から予測されたことを考えると、この分析はうまくいったと思うか。

［うまくいった・うまくいっていない］

演習１０　クラスター分析

クラスター分析の目的

１）類似度による個体（レコード）の分類

２）類似度による変数の分類

クラスター分析は分類をどのように表示するか　→　デンドログラム

デンドログラムの縦軸は　→　要素またはクラスター間の距離（類似の程度を示す量）

要素間の距離とは

個体間について

量的データ：ユークリッド距離、標準化ユークリッド距離、マハラノビス距離等

質的 0/1 データ：類似比、一致係数、φ係数等を使ったもの

変数間について

量的データ：相関係数、順位相関係数等を使ったもの

質的データ：平均平方根一致係数、一致係数、クラメールのＶ等を使ったもの

要素間の距離を知るには　→　距離行列

クラスター構成法

最短距離法（棒状の分布に最適）

最長距離法（クラスターを分離する能力が高い）

他に、群平均法、重心法、メジアン法、ウォード法

クラスター構成過程を表示するには　→　クラスター構成と距離

問題１

多変量演習 9.txt は学生による授業評価のデータであり、レコード（個体）は１つの授業で調べた質問項目（変数）ごとの平均を表している。このデータからクラスター分析を用いて、個体や変数の類似性の特徴を見出したい。以下の質問に答えよ。

１）ユークリッド距離を用いた場合、1 番と 12 番の距離はいくらか。［　　］

２）クラスター構成法を最長距離法、距離測定法をユークリッド距離とする場合、最初にクラスターを構成するのは何番と何番でそれらの距離はいくらか。

個体［　　］番と個体［　　］番で、距離［　　］

３）上の設定で、最初にクラスターとクラスター、またはクラスターと要素の結合になるのはどのようなクラスター（要素）か。それらに含まれる要素を示せ。またその際の距離はいくらか。

クラスター［　　］とクラスター（要素）［　　］　距離［　　］

4）上の設定でクラスター分析を実行し、4つのクラスターに分けたとき、それらのクラスターに含まれる要素（授業の番号）は何か。

［　　］［　　］［　　］［　　］

5）5番が含まれるクラスターと10番が含まれるクラスターの最も大きな特徴は何か。

5番［　　］　10番［　　］

6）距離測定法を標準化ユークリッド距離（各変数を標準化したときのユークリッド距離）に変えた場合、クラスター構成は大きく変わるか。

［変わる・あまり変わらない］　　注）標準化値＝（値－平均値）／標準偏差

7）これにはどんな理由が考えられるか。

各変数の［　　］があまり違わないから。

8）距離測定法をユークリッド距離とし、クラスター構成法を最短距離法に変えるとクラスター構成は大きく変わるか。［変わる・あまり変わらない］

9）ユークリッド距離の場合、その他のクラスター構成法は最長距離法と最短距離法のどちらに近いか。［最長距離法・最短距離法］

各質問についての分類を行いたいが、距離測定法を1－相関係数として以下の問いに答えよ。

10）最長距離法で上の距離測定法を用いる場合、最初にクラスターを構成するのは何と何で、そのときの距離はいくらか。

変数［　　］と変数［　　］で、距離［　　］

11）上の設定でクラスター分析を行い、変数を3つのクラスターに分類する場合、それらのクラスターに含まれる要素（変数）は何か。

［　　］［　　］［　　］

問題2

別紙の性格類似度テストを実施し、誰が近い性格であるか、またどのクラスター構成法が現実的か検討せよ。

演習１１　クラスター分析２

クラスター分析の目的

１）類似度による個体（レコード）の分類

２）類似度による変数の分類

クラスター分析は分類をどのように表示するか　→　デンドログラム

デンドログラムの縦軸は　→　要素またはクラスター間の距離（類似の程度を示す量）

要素間の距離とは

個体間について

量的データ：ユークリッド距離、標準化ユークリッド距離、マハラノビス距離等

質的 0/1 データ：類似比、一致係数、φ係数等を使ったもの

変数間について

量的データ：相関係数、順位相関係数等を使ったもの

質的データ：平均平方根一致係数、一致係数、クラメールのＶ等を使ったもの

要素間の距離を知るには　→　距離行列

クラスター構成法

最短距離法（棒状の分布に最適）

最長距離法（クラスターを分離する能力が高い）

他に、群平均法、重心法、メジアン法、ウォード法

クラスター構成過程を表示するには　→　クラスター構成と距離

問題１

多変量演習 10.txt は生徒のクラス分けに用いる資料で、２学年の成績及び３回の模擬試験の成績のデータである。このデータからクラスター分析を用いて、個体や変数の類似性の特徴を見出したい。以下の質問に答えよ。

１）以下の距離測定法を用いた場合、1 番と 2 番の距離はいくらか。

ユークリッド距離［　　］　標準化ユークリッド距離［　　］

注）どちらの距離測定法を用いるのが良いのか調べるには、すべての変数の正規性が示されれば、実験計画法のところで学んだ Bartlett の検定も利用できる。等分散であればどちらも大差なく、異分散であれば標準化ユークリッド距離が良い。

２）正規性の検定　正規分布と［みなせる・いえない］

３）等分散性の検定　Bartlett の検定を［利用できる・利用できない］

検定確率［　　］　等分散性があると［みなす・いえない］

４）以上より距離測定法は［ユークリッド距離・標準化ユークリッド距離］とする。

以下は距離測定法に上の選択、クラスター構成法に最長距離法を用いて考える。

５）最初にクラスターを構成するのは何番と何番でそれらの距離はいくらか。

個体［　　］番と個体［　　］番で、距離［　　］

６）生徒を３クラス（クラスター）に分けるとするとそれぞれの組に含まれる要素（個人）は何か。［　　］［　　］［　　］

７）2 番が含まれるクラスターと 3 番が含まれるクラスターの最も大きな特徴は何か。

2 番［　　］　3 番［　　］

８）1 番が含まれるクラスターは上のどちらのクラスターに近いか。

［2 番・3 番］の含まれるクラスター

９）クラスター構成法を最短距離法に変えるとクラスター構成は大きく変わるか。

［変わる・あまり変わらない］

10）どちらのクラスター構成法がより分類がはっきりしているか。

［最短距離法・最長距離法］

11）他のクラスター構成法は最長距離法と最短距離法のどちらに近いか。

［最短距離法・最長距離法］

以下では１－相関係数の距離測定法を用いて各変数についての分類を行う。

12）最長距離法で上の距離測定法を用いる場合、最初にクラスターを構成するのは何と何で、そのときの距離はいくらか。

変数［　　］と変数［　　］で、距離［　　］

13）上の設定でクラスター分析を行い、変数を３つのクラスターに分類する。それらの中に含まれる要素（変数）は何か。

［　　］［　　］［　　］

問題２

性格類似度テストを自分で考え、質問票を作れ。

演習１２　正準相関分析

例　正準相関分析 1.txt のデータを用いて、複数の変数間の相関を求める。

正準相関分析の目的　→　複数の変数からなる２つの群の中で特徴的な量を見出し、それらの最大の相関を求める。

どのようにして相関を考えるのか。

$y = a_1$身長$+ a_2$座高
$z = b_1$体重$+ b_2$胸囲　　正準変数の組 y と z が最大の相関を持つよう係数を選ぶ。

y と z の最大の相関とは　→　正準相関係数（変数の組によって複数ある）

係数はどのように表示されるか。　→　正準相関分析で１群係数と２群係数

正準変数 y と z の各データの値を見るには　→　正準変量値

各変数と同じ群の正準変数との関係は　→　正準負荷量（相関係数)、解釈に利用

各変数と違う群の正準変数との関係は　→　交差負荷量（相関係数)、解釈に利用

複数の正準変数の組が得られるが、他の正準変数の組同士の関係は　→　相関係数 0

問題

多変量演習 11.txt について、成績１・２と模試１・２・３に分け、正準相関分析を利用して以下の問いに答えよ。但し、モデルは共分散行列を用いたもので、第１正準相関係数に関して答えよ。

１）成績と模試の第１正準相関係数はいくらか。［　　］

２）成績と模試の正準変数はそれぞれどう表されるか。

成績正準変数＝［　　］成績１＋［　　］成績２

模試正準変数＝［　　］模試１＋［　　］模試２＋［　　］模試３

３）各変数の正準負荷量の値はいくらか。

成績１	成績２	模試１	模試２	模試３

４）各変数の交差負荷量の値はいくらか。

成績１	成績２	模試１	模試２	模試３

５）各正準変数と最も相関のある同じ組の変数は何か。

成績では［成績１・成績２］、模試では［模試１・模試２・模試３］

６）各正準変数と最も相関のある違う組の変数は何か。

模試とは［成績１・成績２］、成績とは［模試１・模試２・模試３］

７）第１正準変量の寄与率はいくつか。
　１群［　　］　２群［　　］
８）1番の人の各正準変数の値を求めよ。
　成績正準変数［　　］　模試正準変数［　　］
９）成績１ 3.5、成績２ 3.8、模試１ 50、模試２ 60、模試３ 70 の人の各正準変量の値を求めよ。　成績正準変量［　　］　模試正準変量［　　］

次に相関行列を用いたモデルに変更して問いに答えよ。
10）成績と模試の正準相関係数の値に違いはあるか。[同じ・違う]
11）成績と模試の正準変数はそれぞれどう表されるか。
　成績正準変数＝［　　］成績１＋［　　］成績２
　模試正準変数＝［　　］模試１＋［　　］模試２＋［　　］模試３
12）正準変量の変数の中で影響力の高いものを見つけるには、共分散行列と相関行列どちらを用いたモデルがよいか。[共分散行列・相関行列]
13）成績と最も関係のある模試は何か。[模試１・模試２・模試３]
14）模試と最も関係のある成績は何か。[成績１・成績２]
15）相関行列を用いたモデルでは２つの正準変数の平均と不偏分散は同じ値になるようになっているが、それはいくらか。平均［　　］不偏分散［　　］
16）この問題の場合第２正準相関係数について考えなくてもよいか。
　[考えなくてもよい・考えるべき]

演習１３　数量化Ⅰ・Ⅱ・Ⅲ類

数量化Ⅰ類（数量化Ⅰ類 1.txt）

各アイテムのカテゴリ名のデータから、各カテゴリが選択されているかどうかの 0/1 データ x_{ij} に変更し、以下の式で目的変数を予測する。

$$Y = a_{11}x_{11} + a_{12}x_{12} + a_{21}x_{21} + a_{22}x_{22} + a_{23}x_{23}$$

（基準化）カテゴリウェイト　→　上式の係数 a_{ij}

予測値と実測値との相関係数　→　重相関係数

予測値は実測値をどれだけ説明しているか　→　寄与率

各アイテムの重要性は　→　相関／偏相関ボタンのウェイト範囲

予測値と実測値の散布図　→　散布図ボタン

数量化Ⅱ類（数量化Ⅱ類 1.txt）

各アイテムのカテゴリ名のデータから、各カテゴリが選択されているかどうかの 0/1 データ x_{ij} に変更し、以下の式でどの群に所属するか予測する。

$$y = a_{11}x_{11} + a_{12}x_{12} + a_{21}x_{21} + a_{22}x_{22} + a_{31}x_{31} + a_{32}x_{32} + a_{33}x_{33}$$

（基準化）カテゴリウェイト　→　上式の係数 a_{ij}

判別方法は　→　判別得点と群別得点平均をみて、どちらの値に近いかで判定する。

数量化Ⅲ類（数量化Ⅲ類 1.txt）

カテゴリに反応したかどうかを表わす 0/1 データ $x_{i\lambda}$（i：カテゴリ，λ：個体）から、カテゴリと個体とで特徴的な量を求める。

カテゴリウェイトで個体得点を求め、個体得点で個体の分類を行なう。

個体ウェイトでカテゴリ得点を求め、カテゴリ得点でカテゴリの分類を行なう。

問題

多変量演習 12.txt は店舗の売り上げや適性を立地、人通り、競合の分類データで予測または判定しようとするものである。

数量化Ⅰ類

１）立地、人通り、競合のカテゴリウェイトを用いた売り上げの予測式を示せ。

予測売り上げ＝［　　］立地１＋［　　］立地２＋［　　］立地３
＋［　　］人通り１＋［　　］人通り２＋［　　］人通り３
＋［　　］競合１＋［　　］競合２＋［　　］

2）基準化カテゴリウェイトを用いた売り上げの予測式を示せ。
予測売り上げ＝［　　］立地１＋［　　］立地２＋［　　］立地３
＋［　　］人通り１＋［　　］人通り２＋［　　］人通り３
＋［　　］競合１＋［　　］競合２＋［　　］

3）予測式は実測値の変動を何%予測できるか。［　　］%

4）立地：２，人通り：２，競合：１の店舗の売り上げをカテゴリウェイトを用いて予測せよ。［　　］

5）上の店舗の売り上げを基準化カテゴリウェイトを用いて予測すると上と同じ値になるか。[なる・ならない]

数量化Ⅱ類

6）カテゴリウェイトを用いた適性を分ける判別関数を示せ。
判別関数＝［　　］立地１＋［　　］立地２＋［　　］立地３
＋［　　］人通り１＋［　　］人通り２＋［　　］人通り３
＋［　　］競合１＋［　　］競合２＋［　　］

7）基準化カテゴリウェイトを用いた適性の判別関数を示せ。
判別関数＝［　　］立地１＋［　　］立地２＋［　　］立地３
＋［　　］人通り１＋［　　］人通り２＋［　　］人通り３
＋［　　］競合１＋［　　］競合２＋［　　］

8）カテゴリウェイトを用いた適性の判別関数を使って、問題４）の店舗についての判別得点を求めよ。［　　］

9）基準化カテゴリウェイトを用いた適性の判別関数を使って、判別得点が上と同じ値になるか。[なる・ならない]

10）この店舗の適性はどちらと思われるか。[１・２]

11）数量化理論では売り上げの予測値や判別関数の値はとびとびの値が得られる。このデータでは何種類の値が得られるか。［　　］種類

演習１４　時系列分析

1. 変動の分解モデル

時系列データを傾向変動、季節変動、循環変動、残差に分解し、データの性質を調べると同時に予測も行う手法で、データに周期性がある場合に有効

　傾向変動　全体的な変動の傾向を表す変動

　季節変動　一定の周期を持つ変動

　循環変動　一定の周期ではない変動（ここでは長期の周期変動を考えている）

　残差　　これらの変動を差し引いた残りの変動

時系列データを見る　→　「元データ」ラジオボックスを選択し、描画ボタン

傾向変動を分解する　→　近似モデルで見てよく適合するモデルを求める。

変動の分解の表示で、元データ、傾向変動、残差をチェックし、実行

周期性を見る　→　コレログラム（自己相関のグラフ）とピリオドグラムで調べる。

季節（循環）変動を分解する　→　分解の周期を入力し、表示にその変動を加え実行

どの程度予測があっているかの目安　→　残差 2 乗平均の値、R^2 値

2. 予測モデル

歴史的モデル

　差の平均法、指数平滑法、ブラウン法（2 重指数平滑法)、移動平均法

　傾向変動が大きい場合に利用する。パラメータを含むものもある。

比較的最近のモデル

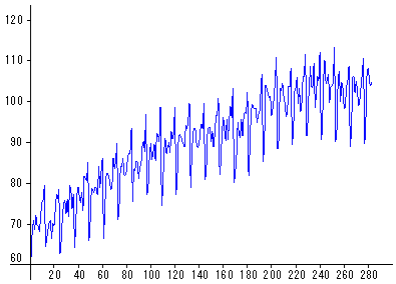
　最近隣法、ARIMA(p, r, q)

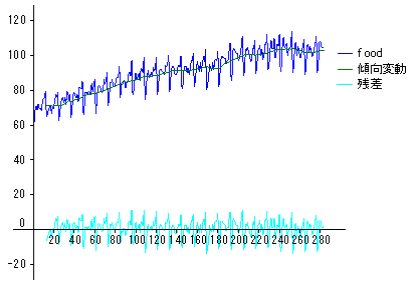
　傾向変動がある場合は除いて利用する。

問題１

　Samples¥時系列（decomp）.txt の food について、以下の問いに答えよ。

１）このデータをグラフで表せ。（左下）





2）傾向変動を 12 期の移動平均で推定した場合のグラフを示せ。(右上)

3）傾向変動を除いた残差から、コレログラムを用いて季節変動の周期を求めるといくらか。(ヒント：これは 1 ヶ月単位のデータ)　［　　］

4）上の季節変動を除いた場合の残差 2 乗平均の値はいくらか。　［　　］

5）ピリオドグラムのデータより、周期 45 から 55 の間で残差 2 乗平均の値を最小にする循環変動の周期はいくらか。　［　　］

6）上の循環変動を除いた場合の残差 2 乗平均の値はいくらか。　［　　］

7）データを上の傾向変動、季節変動、循環変動で予測するモデルの R^2 の値はいくらか。　［　　］

8）このモデルでの 1 期先の予測値はいくらか。　［　　］

予測モデル

9）以下のモデルで、最適な実測・予測の R^2 値、残差 2 乗平均と 1 期先の予測値を求めよ。

	実測・予測 R^2	残差 2 乗平均	次期予測値
差の平均法			
指数平滑法			
ブラウン法			

10）最近隣法で２）の傾向変動の分解を行わない場合と上の方法で行う場合の残差 2 乗平均を求めよ。　行わない場合［　　］，行う場合［　　］

11）上の方法で傾向変動の分解を行った場合の R^2 値を求めよ。
［　　］

12）２）の傾向変動の分解を行った ARIMA モデルで、残差の 2 乗平均を最小にする最適なパラメータとそのときの残差 2 乗平均を求めよ。但し $p=12$ とする。
（p，d，q）＝（１２，　，　）　残差 2 乗平均［　　］

13）変動の分解モデルと予測モデルではどちらが有効と思えるか。
［分解モデル・予測モデル］

14）このデータの 12 期先を予測したい。傾向変動の除去は次のどれをつかうべきか。
［12 期移動平均・べき乗近似・2 次多項式近似］

問題２

Samples¥時系列（decomp）.txt のデータ sunspot について、傾向変動を 3 期の移動平均として変動の分解モデルを用いて 1 期先の値を予測せよ。