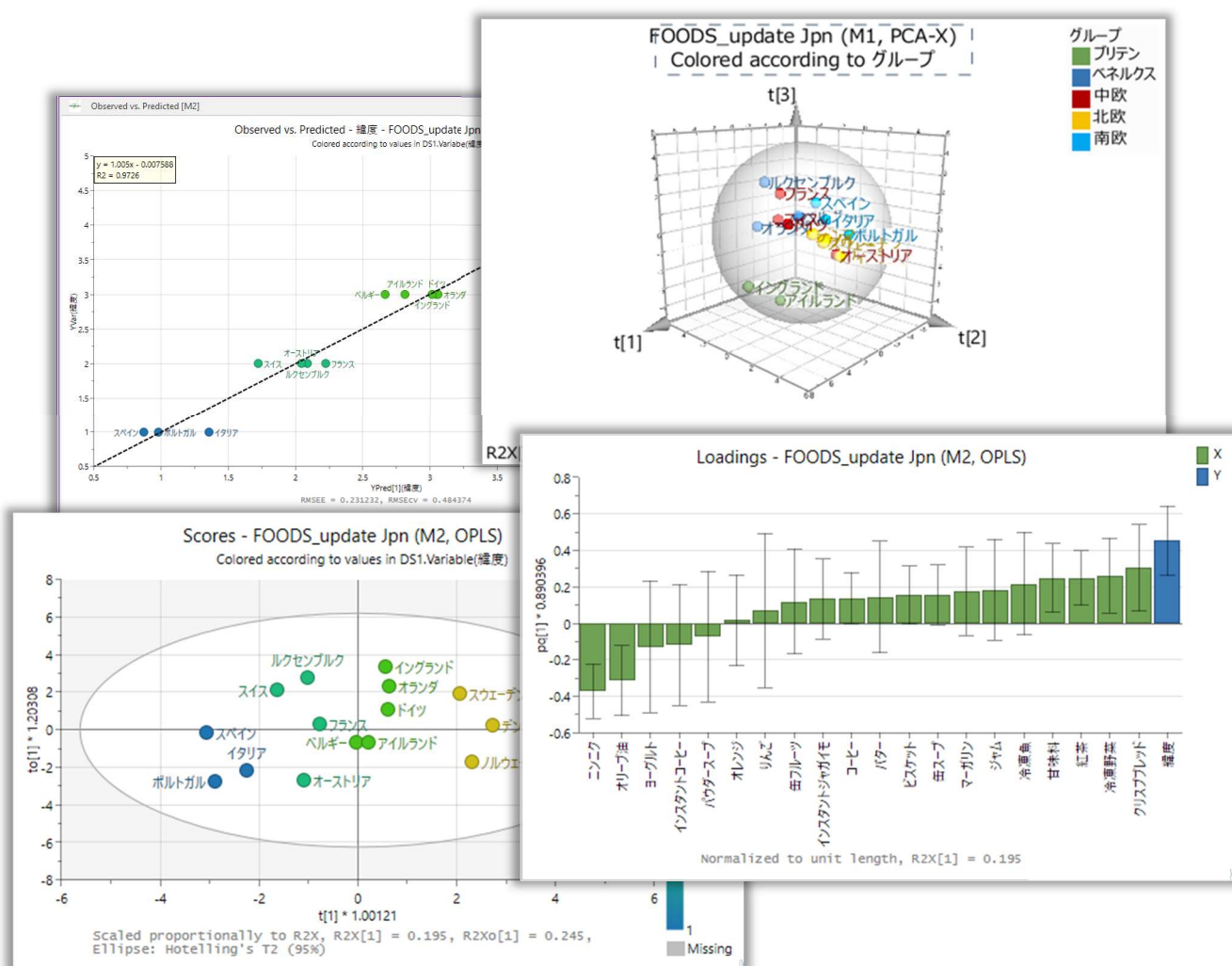


## SIMCA18 チュートリアル

### ～Foods データ(PCA,OPLS)編～



## 目次

1.データセット.....	1
2.主成分分析（PCA） .....	6
モデルの作成 .....	6
結果の解釈.....	11
3.OPLS .....	19
モデルの作成 .....	19
結果の解釈.....	27
回帰モデルを用いた予測.....	30

## 備考：

本チュートリアルは、SIMCA version 18.0.0 に基づいて作成しております。そのため結果の一部（スコア、可視化部分も含む）や画面がお客様と異なる場合がございます。ご了承ください。

# 1. データセット

モデル構築のためのデータを読み込みます。利用するデータにはヨーロッパ諸国の（16 か国）の食品消費量と緯度情報が含まれています。

## データセットの準備

EXCEL 形式で用意します。

ファイル名：「FOODS\_update Jpn.xlsx」

食品消費データと緯度

- ・ サンプル：ヨーロッパ 16 か国
- ・ 変数：食品 20 種（保存食）、首都の緯度（5 段階）
- ・ 目的： 1）食品データの概観（PCA）  
2）緯度に関する食品を調べる（OPLS）

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW DISPLAY HELP

FOODS\_update 1pn.xlsx Excel 検索 履歴 表示 コメント

実行したい作業を入力してください

Meiryo UI 10 A A

セルの範囲を選択して全体的に表示する

セルを結合して中央揃え

条件付き書式

スタイル

挿入

削除

書式

ΣオートSUM

ファイル

クリア

並べ替え

フィルター

検索とナビ

クリップボード

フォント

配置

数値

スタイル

セル

編集

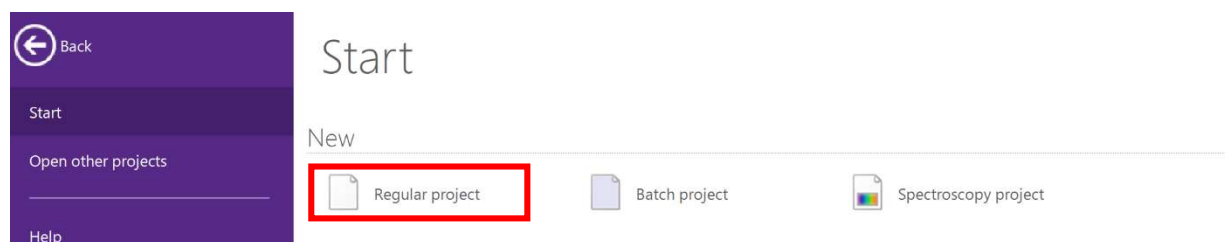
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	国	グループ	緯度	ヨーグルト	パン	鶏肉	豚肉	牛肉	魚	野菜	果物	豆	卵	乳製品	穀類	油脂	調味料	飲料	その他	合計				
2	ドイツ	中欧	3	90	49	88	19	57	51	19	21	27	21	81	75	44	71	22	91	85	74	30	26	
3	イタリア	南欧	1	82	10	60	2	55	41	3	2	4	2	67	71	9	46	80	66	24	94	5	18	
4	フランス	中欧	2	88	42	63	4	76	53	11	23	11	5	87	84	40	45	88	94	47	36	57	3	
5	オランダ	北欧	3	96	62	98	32	62	67	43	7	14	14	83	89	61	81	15	31	97	13	53	15	
6	ベルギー	北欧	3	94	38	48	11	74	37	23	9	13	12	76	76	42	57	29	84	80	83	20	5	
7	ルクセンブルク	北欧	2	97	61	86	28	79	73	12	7	26	23	85	94	83	20	91	94	94	84	31	24	
8	イングランド	中欧	3	27	86	99	22	91	55	76	17	20	24	76	68	89	91	11	95	94	57	11	28	
9	ポルトガル	南欧	1	72	26	77	2	22	34	1	5	20	3	22	51	8	16	89	65	78	92	6	9	
10	オーストリア	中欧	2	55	31	61	15	29	33	1	5	15	11	49	42	14	41	51	51	72	28	13	11	
11	スイス	中欧	2	73	72	85	25	31	69	10	17	19	15	79	70	46	61	64	82	48	61	48	30	
12	スウェーデン	北欧	4	97	13	93	31	43	43	39	54	45	56	78	53	75	9	68	32	48	2	93		
13	デンマーク	北欧	4	96	17	92	35	66	32	17	11	51	42	81	72	50	64	11	92	91	30	11	34	
14	ルクセンブルク	北欧	4	92	17	83	13	62	51	4	17	30	15	61	72	34	51	11	63	94	28	2	62	
15	フィンランド	北欧	5	90	12	04	20	64	27	10	0	10	12	50	57	22	37	15	96	94	17	64		
16	スペイン	南欧	1	70	40	40	20	62	43	2	14	23	7	59	77	30	38	86	44	51	91	16	13	
17	アイスランド	北欧	3	30	52	99	11	80	75	18	2	5	3	57	52	46	89	5	97	25	31	3	9	
18																								
19																								
20																								
21																								
22																								
23																								
24																								
25																								
26																								
27																								
28																								
29																								
30																								
31																								
32																								
33																								
34																								
35																								
36																								
37																								
38																								
39																								
40																								
41																								

Sheet1

## データセットのインポート

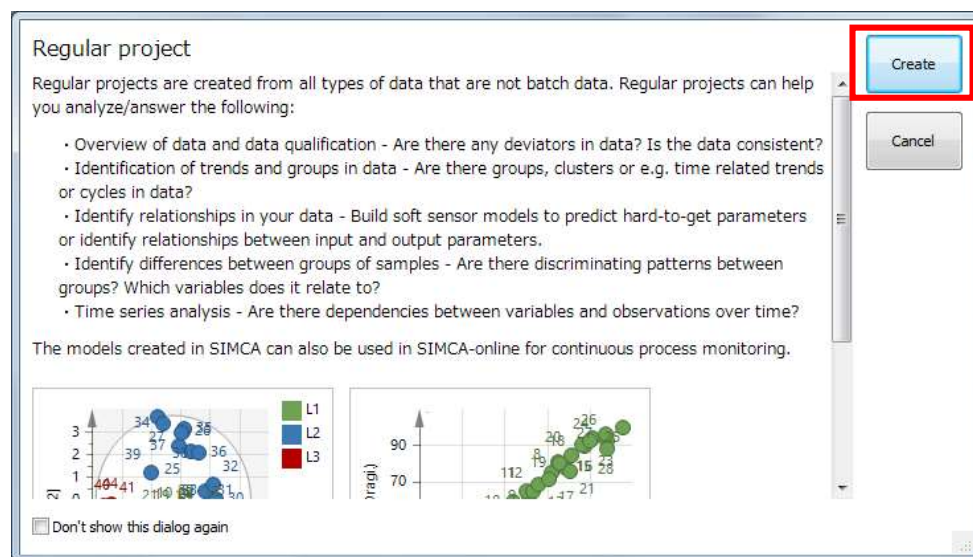
SIMCA18 を起動します。

Start 画面が立ち上がるので、Regular project をクリックします。SIMCA ではプロジェクト単位でデータの読み込みとモデルの構築を行います。

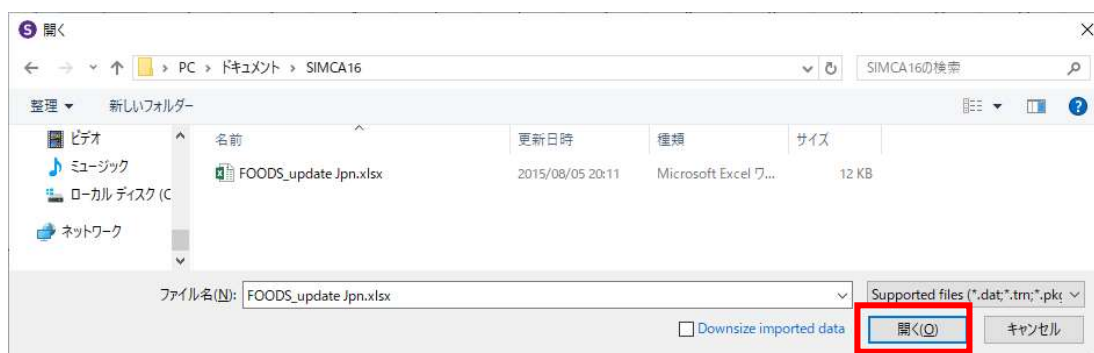


💡 3 種プロジェクトが存在しますが、選択に合わせてよく使う機能のメニュー表示のみが変わり（リボンメニュー機能）、基本的には差異はありません。よって Regular project の利用を推奨いたします。

Create をクリックします。



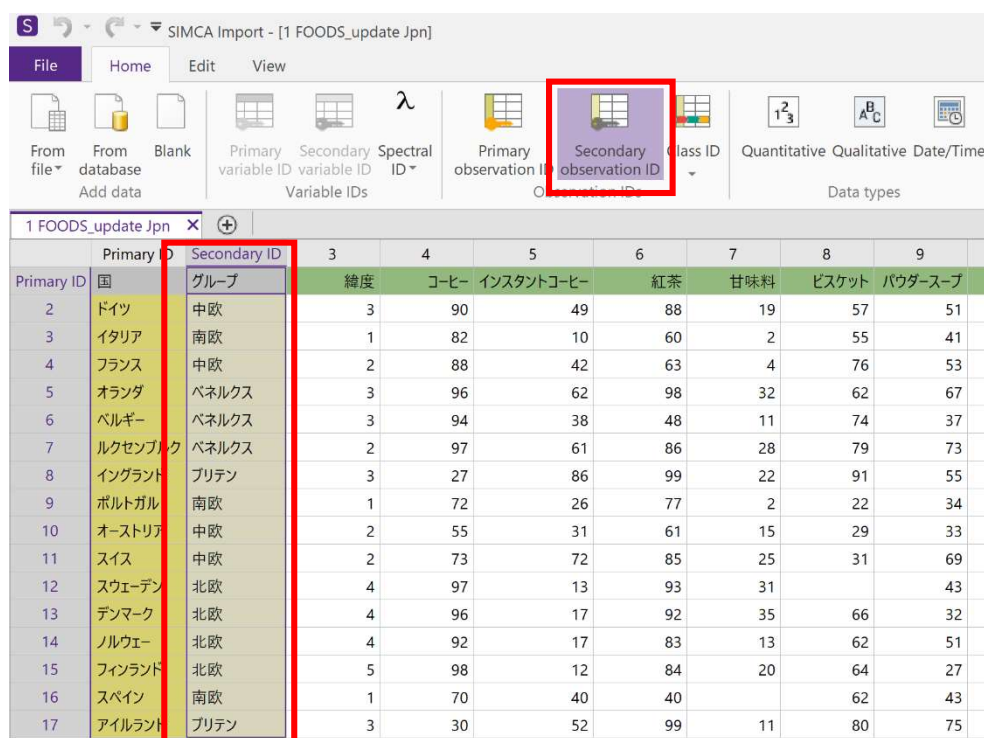
用意していたデータセットを指定して、開きます。



SIMCA は読み込んだデータセットに対し、自動でデータの属性を設定します。ユーザー側で指定する場合には設定を変更する必要があります。

第 2 カラム（グループ）をクリックし、列全体を選択します。

Home > Secondary observation ID をクリックすると、第 2 カラムが Secondary ID に設定されます。



💡 縦・横ともに、Primary ID を 1 つずつ・ユニークなものが必要になります。

第3カラム（緯度）はOPLS解析のY変数に値するので、先ほどと同様の手順でY-variableに変更します。

Primary ID	Secondary ID	Spectral ID	緯度	4	5	6	7	8	9	10	11	12	13	14	15	16	17
2	ドイツ	中欧	3	90	49	88	19	57	51	19	21	27	21	81	75	44	71
3	イタリア	南欧	1	82	10	60	2	55	41	3	2	4	2	67	71	9	46
4	フランス	中欧	2	88	42	63	4	76	53	11	23	11	5	87	84	40	45
5	オランダ	ベネルクス	3	96	62	98	32	62	67	43	7	14	14	83	89	61	81
6	ベルギー	ベネルクス	3	94	38	48	11	74	37	23	9	13	12	76	76	42	57
7	ルクセンブルク	ベネルクス	2	97	61	86	28	79	73	12	7	26	23	85	94	83	20
8	イングランド	ブリテン	3	27	86	99	22	91	55	76	17	20	24	76	68	89	91
9	ポルトガル	南欧	1	72	26	77	2	22	34	1	5	20	3	22	51	8	16
10	オーストリア	中欧	2	55	31	61	15	29	33	1	5	15	11	49	42	14	41
11	スイス	中欧	2	73	72	85	25	31	69	10	17	19	15	79	70	46	61
12	スウェーデン	北欧	4	97	13	93	31		43	43	39	54	45	56	78	53	75
13	デンマーク	北欧	4	96	17	92	35	66	32	17	11	51	42	81	72	50	64
14	ノルウェー	北欧	4	92	17	83	13	62	51	4	17	30	15	61	72	34	51
15	フィンランド	北欧	5	98	12	84	20	64	27	10	8	18	12	50	57	22	37
16	スペイン	南欧	1	70	40	40		62	43	2	14	23	7	59	77	30	38
17	アイルランド	ブリテン	3	30	52	99	11	80	75	18	2	5	3	57	52	46	89

残りのカラムはデータ（X変数）であり変更はありません。

データセットの設定が完了したので“Finish import”をクリックし、インポートを終了します。

ID ならびに変数指定に問題がなければ画面左下に「No issues」が表示されます。問題がある場合は、画面左下に warning が表示されますのでクリックしていただき問題箇所（Go to）をご確認ください。

16	スペイン	南欧	1	70	40	40		62	43	2
17	アイルランド	ブリテン	3	XX	52	99	11	80	75	18

Issues

Issue

Sheet: 1 FOODS\_update Jpn (21 variables, 16 observations, 4 (1.19 %) missing)

Go to → Invalid value

Resolve all

1 warning



ファイル名、場所を指定して、読み込み完了です。



## データセットの確認

Home > Datasetをクリックして、インポートしたデータセットを表示します。

欠損値（Missing values）がピンクのセルで表されますが、SIMCA では欠損値を考慮した計算が可能ですので、問題ありません。

FileHomeDataAnalyzePredictViewToolsDeveloperAdd-ins

ProjectDataset

New as Edit Delete

Statistics Compare models

Model type

Two first Add Remove

Summary of fit

Overview

Scores Loadings Hotelling's T<sup>2</sup> DMod P<sub>T</sub>

Obs. vs. pred. Coefficients VIP

Create Plot/list

WorksetModelDiagnostics & Interpretation

1 FOODS\_update.jpnnSIMCA - [1 FOODS\_update.jpnn

Active model: N/A

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	
Primary ID	グループ	国産	ユービー	インスタントコーヒー	紅茶	甘味料	ビスケット	パウダースープ	缶スープ	インスタントジャガイモ	冷凍魚	冷凍野菜	りんご	オレンジ	加工フルーツ	ジャム	ニンニク	バター	
2	ドイツ	中欧	3	90	49	88	19	57	51	19	21	27	21	81	75	44	71	22	91
3	イタリア	南欧	1	82	10	60	2	55	41	3	2	4	2	67	71	9	46	80	66
4	フランス	中欧	2	88	42	63	4	76	53	11	23	11	5	87	84	40	45	88	94
5	オランダ	パネルクス	3	96	62	98	32	62	67	43	7	14	14	83	89	61	81	15	31
6	ベルギー	パネルクス	3	94	38	48	11	74	37	23	9	13	12	76	76	42	57	29	84
7	ルクセンブルク	パネルクス	2	97	61	86	28	79	73	12	7	26	23	85	94	83	20	91	94
8	イングランド	ブリタニ	3	27	86	99	22	91	55	76	17	20	24	76	68	89	91	11	95
9	ポルトガル	南欧	1	72	26	77	2	22	34	1	5	20	3	22	51	8	16	89	65
10	オーストラリア	中欧	2	55	31	61	15	29	33	1	5	15	11	49	42	14	41	51	51
11	スイス	中欧	2	73	72	85	25	31	69	10	17	19	15	79	70	46	61	64	82
12	スウェーデン	北欧	4	97	13	93	31	43	43	39	54	45	56	78	53	75	9	68	
13	デンマーク	北欧	4	96	17	92	35	66	32	17	11	51	42	81	72	50	64	11	92
14	ノルウェー	北欧	4	92	17	83	13	62	51	4	17	30	15	61	72	34	51	11	63
15	フィンランド	北欧	5	98	12	64	20	64	27	10	8	18	12	50	57	22	37	15	96
16	スペイン	南欧	1	70	40	40	40	62	43	2	14	23	7	59	77	30	38	86	44
17	アイルランド	ブリタニ	3	30	52	99	11	80	75	18	2	5	3	57	52	46	89	5	97

Quick Info | 1 FOODS\_update.jpnn

Variable | Observation | Dataset

Missing values

Statistics

Variables

X-variables

Y-variables

Chocography

Missing values: 3 (0.8%)

💡 SIMCA では欠損値とゼロを分けて認識します。空欄は欠損値とみなされます。

以上で、データセットのインポートが完了となります。

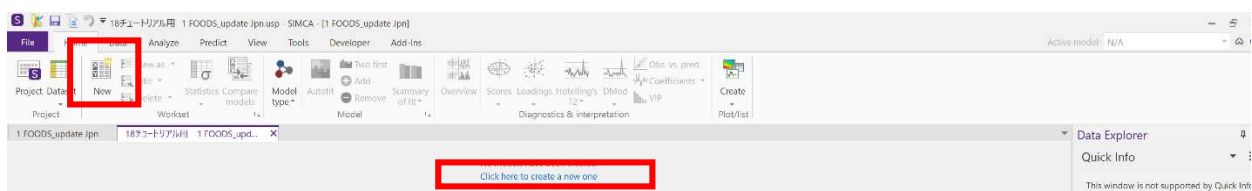
## 2.主成分分析(PCA)

主成分分析により、データの外観を捉えます。

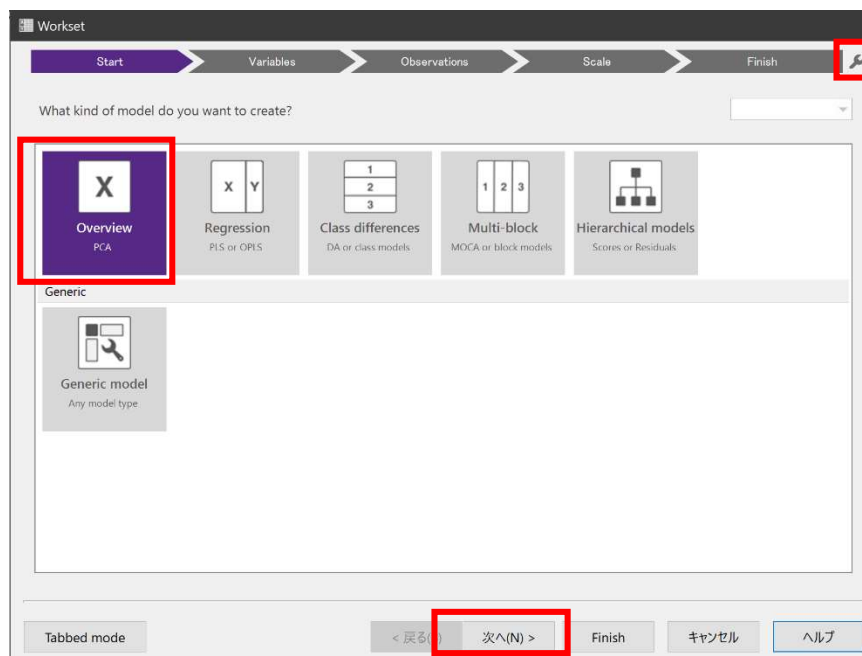
### モデルの作成

#### モデルを設定する

Home > New または “Click here to create a new one” をクリックします。



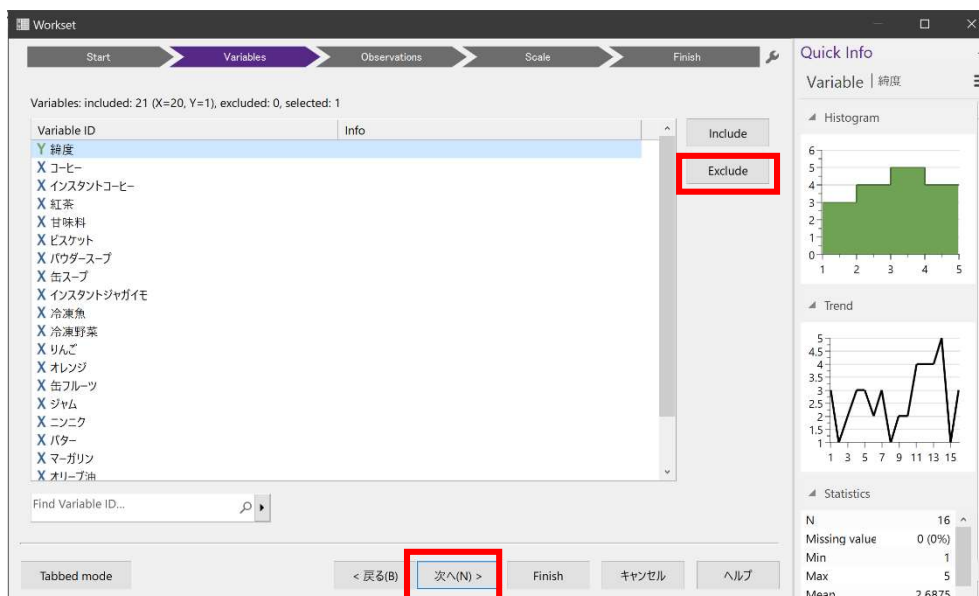
Overview (PCA) を選択し、次へをクリックします。



上図のように Start、Variables、Observations、Scale、Finish が表示されない場合はスパナのアイコンで設定します。

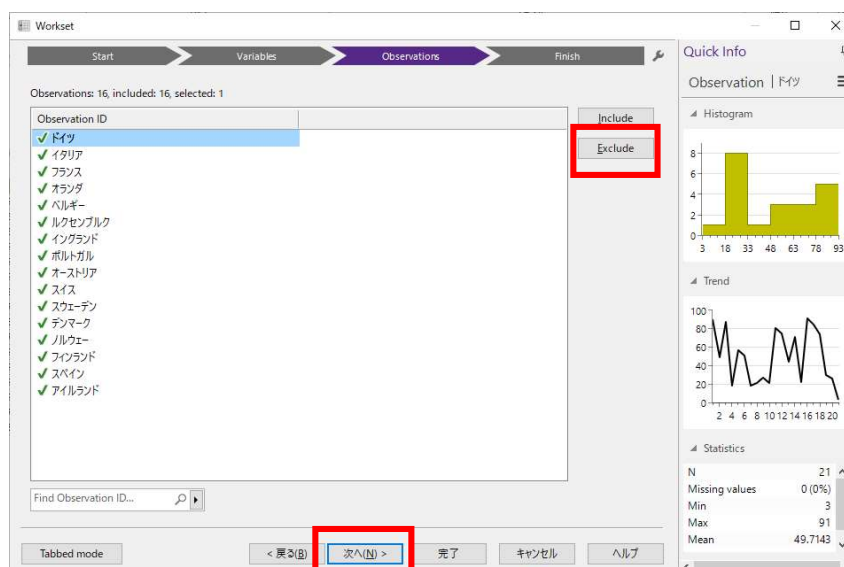


Variables（変数）の設定をします。Xは説明変数、入力、原因のとなり、Yは目的変数、出力、結果となります。Y変数は回帰（Regression）モデルである PLS、OPLS で利用します。不要な変数があれば Exclude をクリックして除外します。



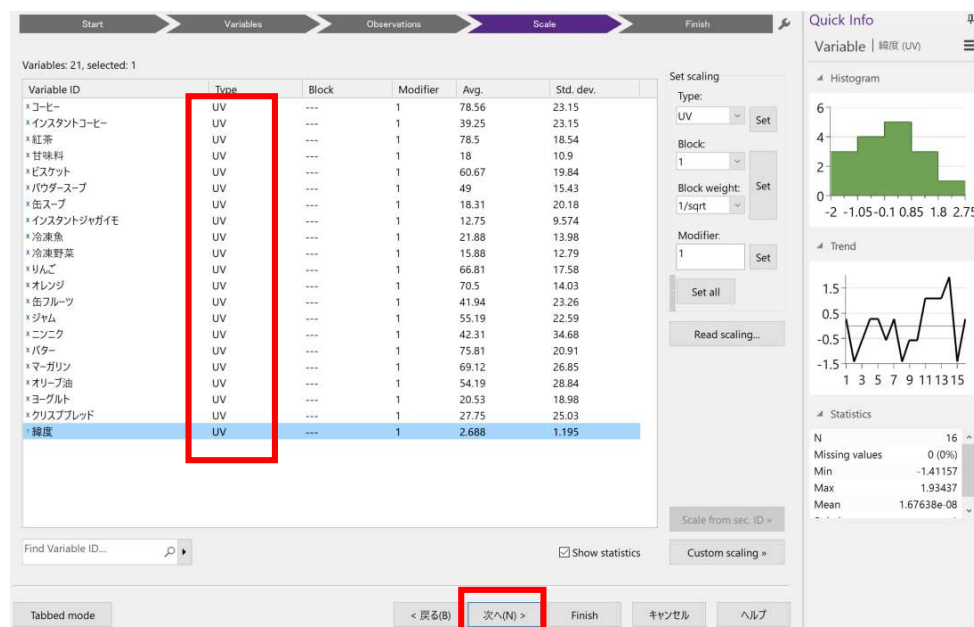
ここでは何も操作をせず、次へをクリックします。

Observations（観測サンプル）の設定をします。不要なサンプルがあれば Exclude をクリックして除外します。



ここでも何も操作せず、次へをクリックします。

Scale（スケーリング：尺度を合わせる）の設定をします。不要なサンプルがあれば Type が UV になっていることを確認します（スケーリングは、MS メタボロミクスデータを使った解析編で説明いたします）



Variables: 21, selected: 1

Variable ID	Type	Block	Modifier	Avg.	Std. dev.
* コーヒー	UV	---	1	78.56	23.15
* インスタントコーヒー	UV	---	1	39.25	23.15
* 紅茶	UV	---	1	78.5	18.54
* 甘味料	UV	---	1	18	10.9
* ビスケット	UV	---	1	60.67	19.84
* パウダースープ	UV	---	1	49	15.43
* 缶スープ	UV	---	1	18.31	20.18
* インスタントジャガイモ	UV	---	1	12.75	9.574
* 冷凍魚	UV	---	1	21.88	13.98
* 冷凍野菜	UV	---	1	15.88	12.79
* リンゴ	UV	---	1	66.81	17.58
* オレンジ	UV	---	1	70.5	14.03
* 缶フルーツ	UV	---	1	41.94	23.26
* ジャム	UV	---	1	55.19	22.59
* ニンニク	UV	---	1	42.31	34.68
* バター	UV	---	1	75.81	20.91
* マーガリン	UV	---	1	69.12	26.85
* オリーブ油	UV	---	1	54.19	28.84
* ヨーグルト	UV	---	1	20.53	18.98
* クリスプブレッド	UV	---	1	27.75	25.03
緯度	UV	---	1	2.688	1.195

Set scaling  
Type: UV Set  
Block: 1 Set  
Block weight: 1/sqrt Set  
Modifier: 1 Set  
Set all  
Read scaling...

Quick Info  
Variable | 緯度 (UV)  
Histogram  
Trend  
Statistics  
N 16  
Missing values 0 (0%)  
Min -1.41157  
Max 1.93437  
Mean 1.67638e-08

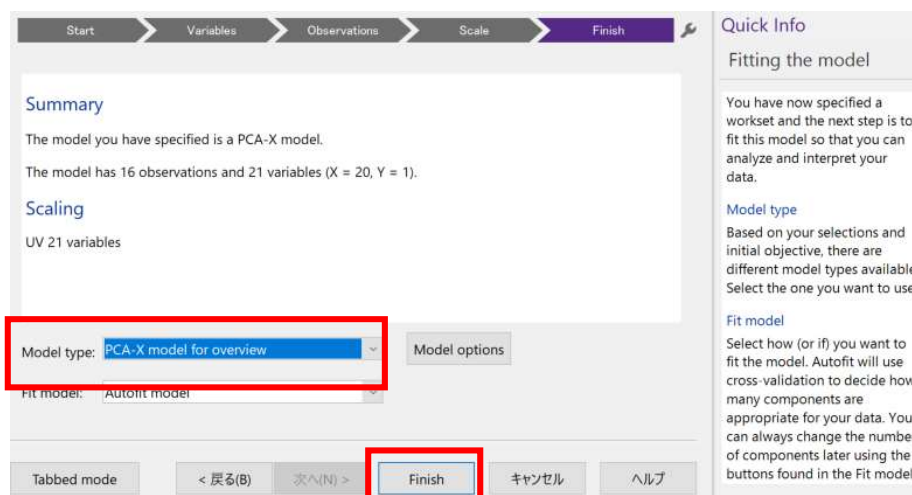
Find Variable ID... Show statistics Custom scaling >

Tabbed mode < 戻る(B) 次へ(N) > Finish キャンセル ヘルプ

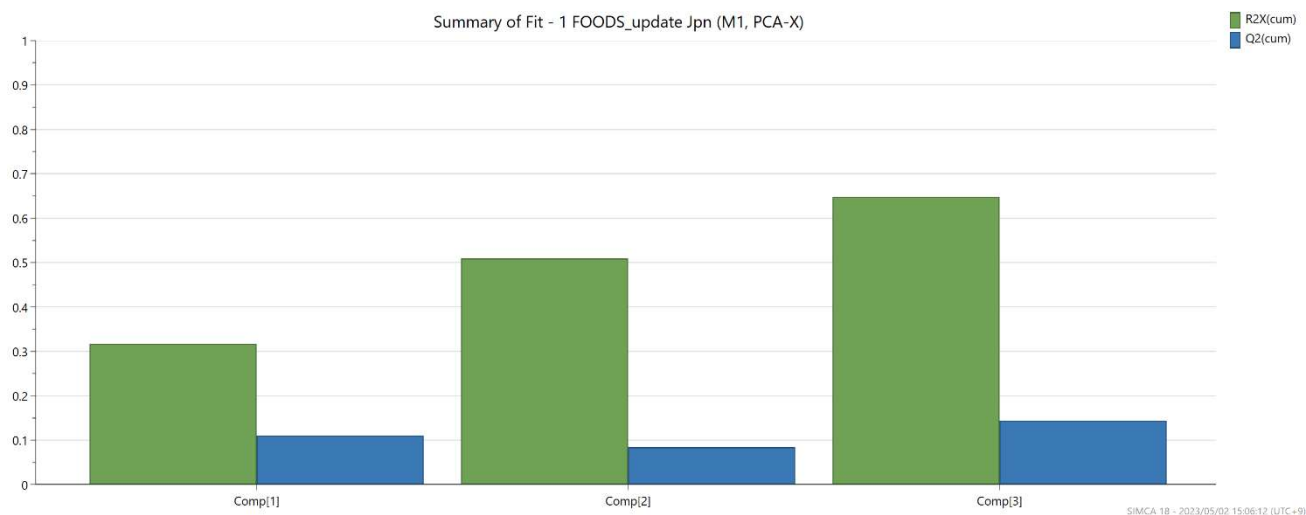
次へをクリックします。

Summary 画面ではモデルタイプ (model type)、変数 (variables)、観測値 (observations)、スケーリング (scaling) が表示され、最終確認を行います。

Model types を “PCA-X model for overview” に設定し、Finish をクリックします。



SIMCA が自動的に最適なモデルを計算し、モデルを作成します。



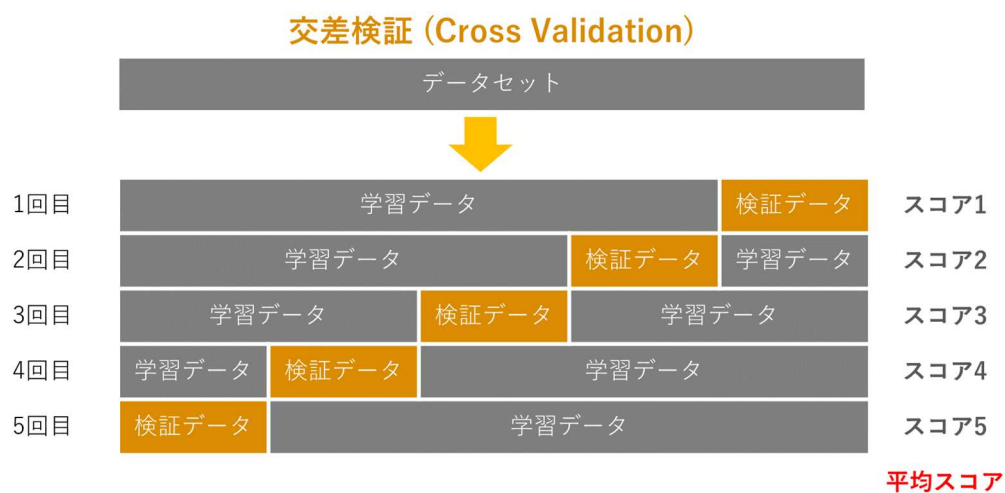
ここでは、第 3 主成分まで計算されました。

緑はその成分でどれだけのデータを表しているかの寄与率(R2)、青は交差検定による寄与率(Q2)を表します。



R2：寄与率でデータの当てはまりの良さを示します。総変動量の中で説明できた変動の割合をしめし、0.0～1.0で示されます。1.0 が完全な当てはまりになります。

Q2：交差検証による寄与率を示します。交差検証とはデータの一部を除外してモデルを作成し、除外したデータを検証データとして評価します。これを複数回行い、平均したものを Q2 として算出します。



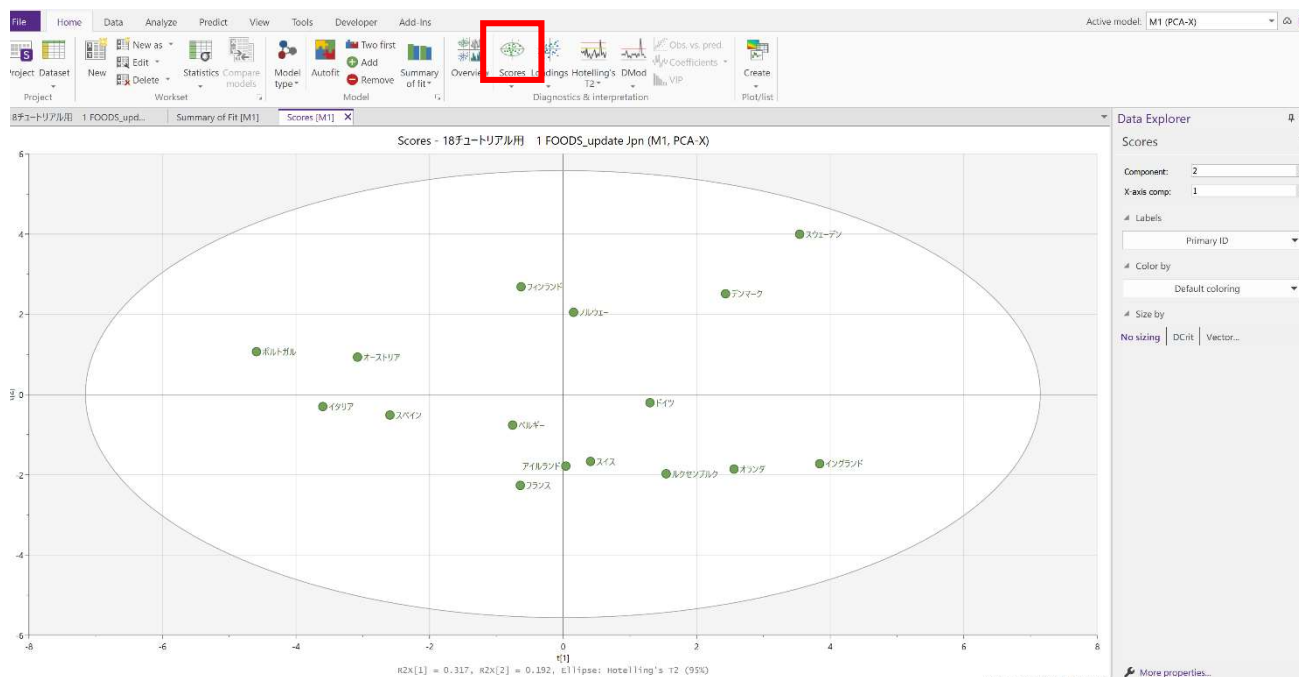
常に  $R2 > Q2$  となります。両スコアがともに増加するモデルが良好とされます。

## 結果の解釈

スコアプロットとローディングプロットを中心に各種プロットを表示し、PCA 結果の解釈・点検を行います。

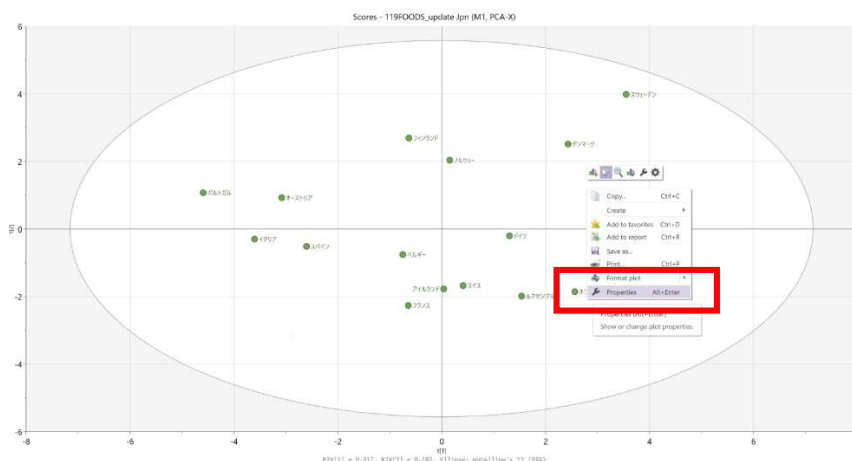
### スコアプロット

Home>Scores をクリックし、スコアプロットを表示させます。



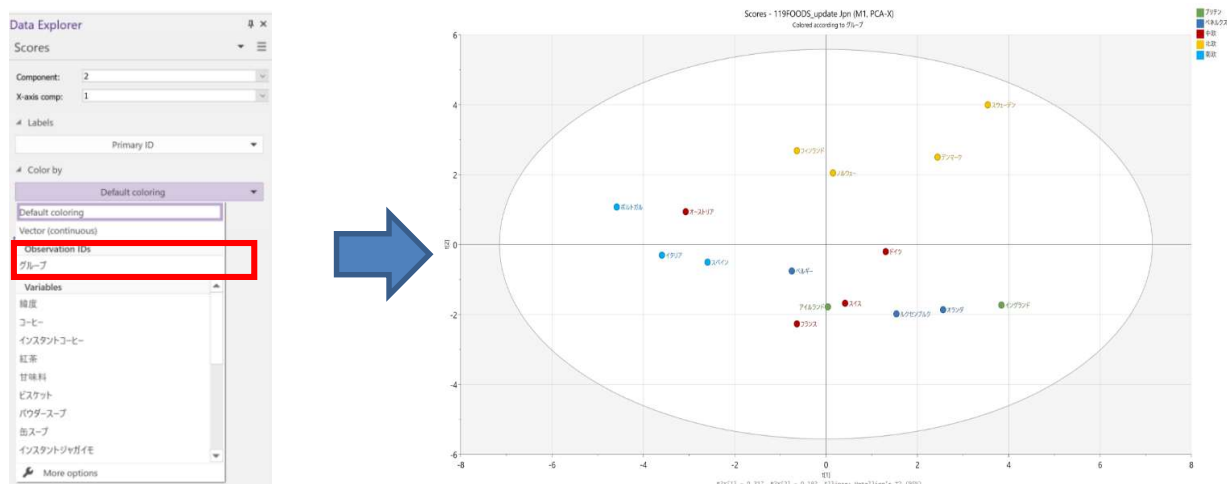
スコアプロットでは、横軸  $t[1]$  がデータのばらつきが最大となるように求められた第 1 主成分、縦軸  $t[2]$  が第 2 主成分として、デフォルトでは設定されています。

サンプルをクラス(ここではグループ)ごとに色付けをします。画面右端に Data Explorer が表示されてない場合はスコアプロット上で右クリックし Properties を選択します。



Copyright Infocom Corporation 本資料の無断複製および転用は禁じられています

Data Explorer > Color by から、グループを選択すると、スコアプロットでサンプルが色付けされます。

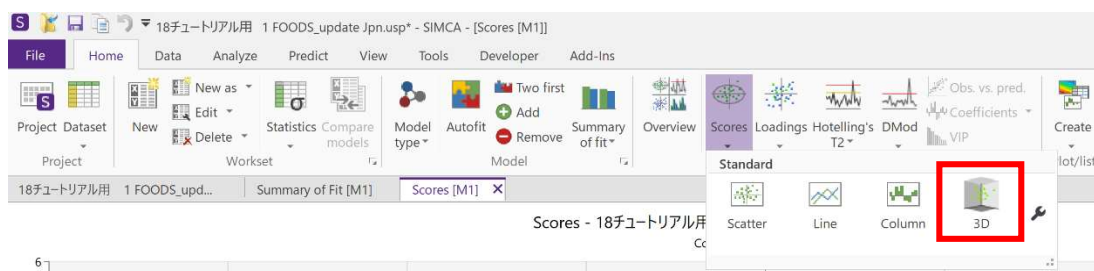


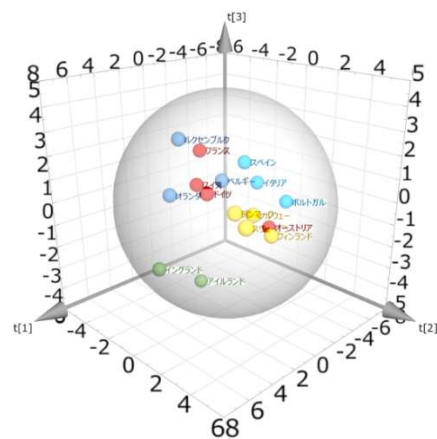
グループごと（ここでは地域ごとに）に似た傾向があることが確認できます。

## 立体的に分布を見る

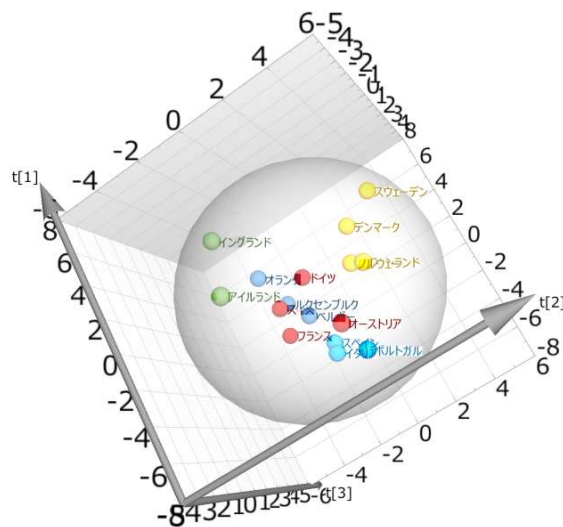
SIMCA では 3D での表示することで立体的にサンプルやピークの分布を調べることが可能です。

Home > Scores > 3D をクリックしてください。





マウスをドラッグすることでプロットを回転させたり、ドラッグ・アンド・ドロップすることで自動的に回転させたりすることができます。(右クリック → Reset rotation で元の角度に戻せます。)



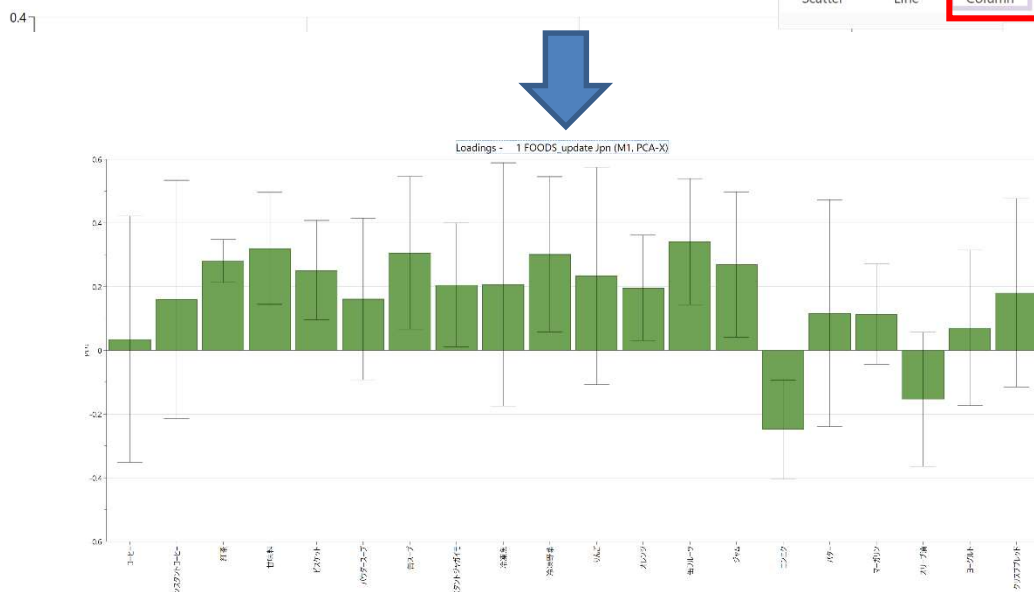
グループで似た傾向をとっていることを、より視覚的に捉えることができます。

## ローディングプロット

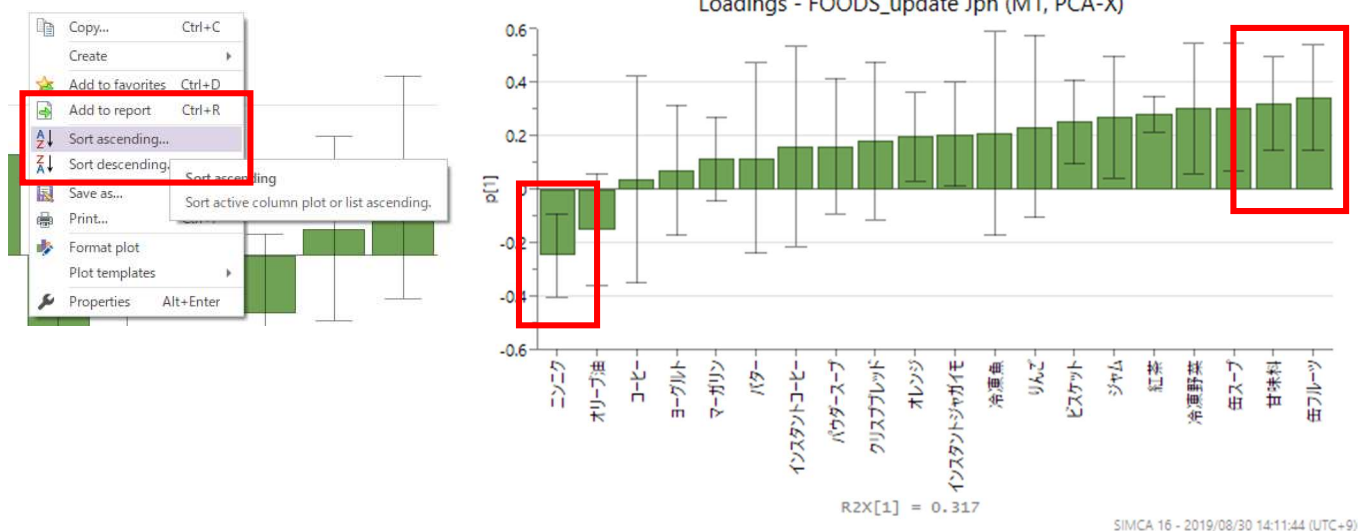
ローディングプロットは主成分と元の変数との相関を表します。Home>Loadings をクリックし、ローディングプロットを表示させます。



The screenshot shows the SIMCA software interface. The top menu bar includes 'File', 'Home', 'Data', 'Analyze', 'Predict', 'View', 'Tools', 'Developer', and 'Add-Ins'. The 'Analyze' menu is open, showing options like 'New as', 'Edit', 'Delete', 'Statistics', 'Compare models', 'Model type', 'Autofit', 'Remove', 'Summary of fit', 'Overview', 'Scores', 'Loadings', 'Hotelling's T2', 'DMod', 'VIP', and 'Create'. The 'Loadings' menu is highlighted, and the 'Column' plot type is selected, indicated by a red box around the 'Column' icon in the bottom right corner.



表示されたカラムプロット上で右クリックし、Sort ascending で大きい順に並べ替えます。

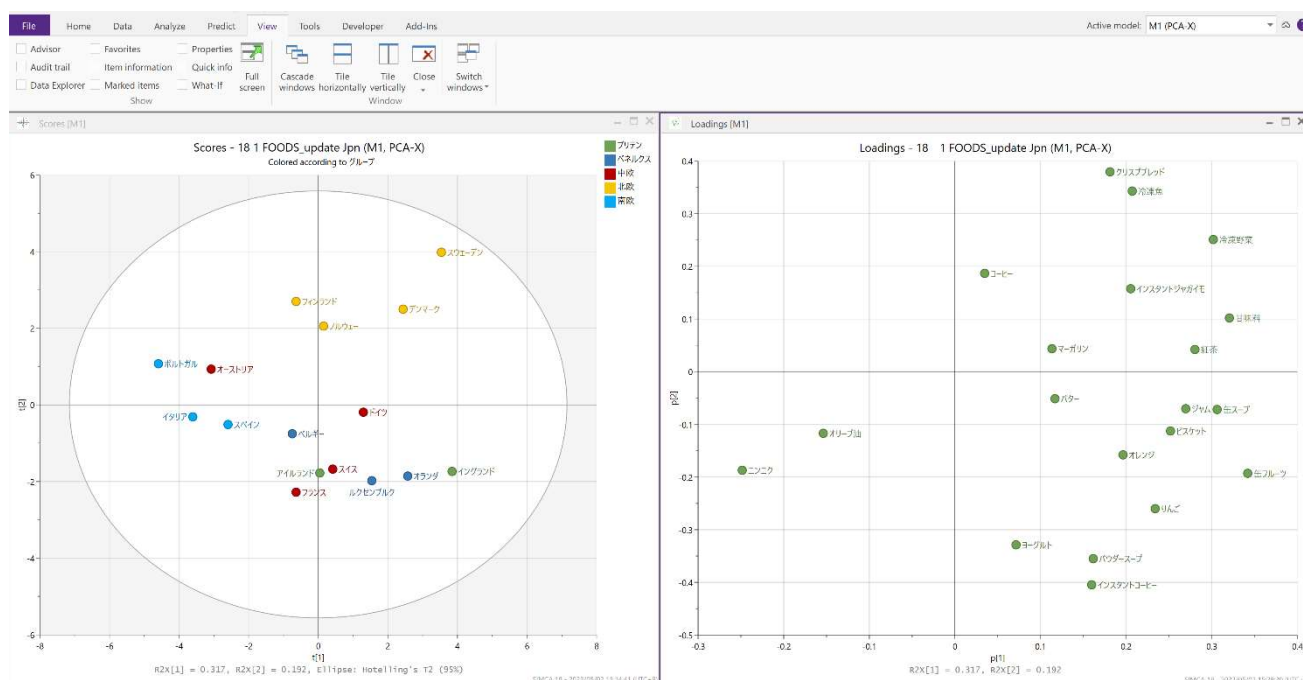


第 1 成分 (p[1]) では、缶フルーツや甘味料が正の相関が高く、にんにくが負の相関が高いことがわかります。

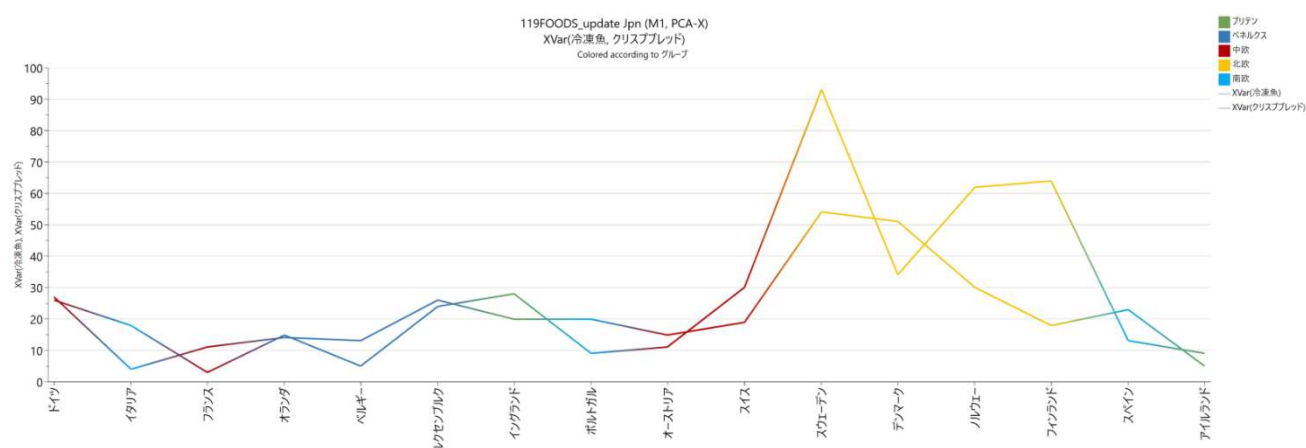
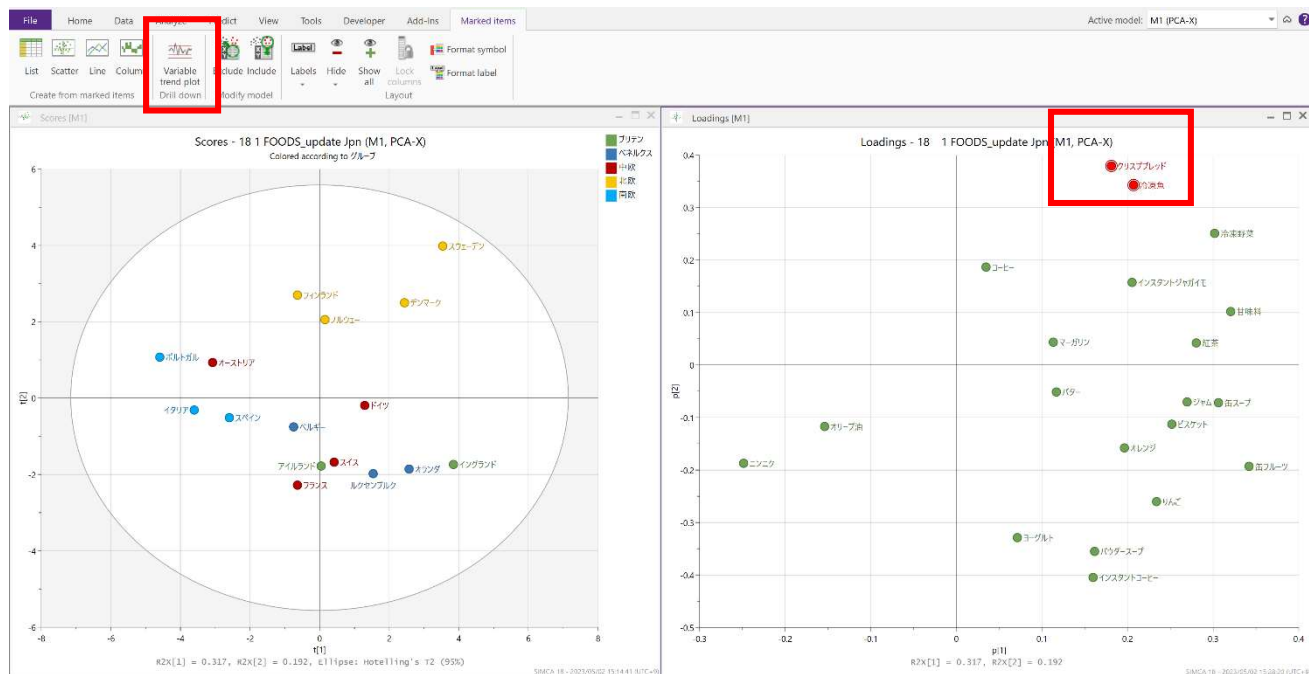
💡 上記例におけるにんにくは負の相関であって相関がないわけではないことにご注意ください。

## 特定ポイント(変数)のトレンドを見る

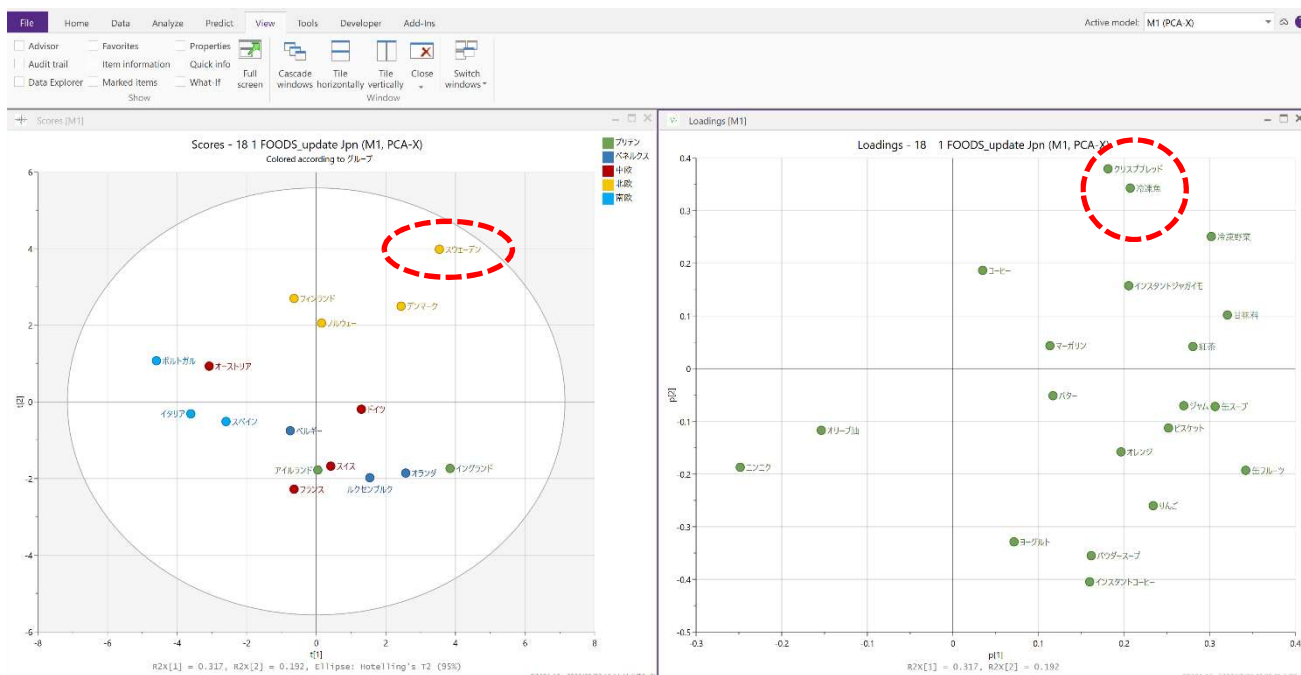
変数の詳細を確認することで解析結果の解釈を深めます。スコアプロットとローディングプロットを並べて表示させます。



ローディングプロット上で、クリスマスブレッドと冷凍魚を、マウスのドラッグで選択します。Marked items > Variable trend plot をクリックすると、Xvar Plot が表示され、選択した変数のトレンドを見ることが出来ます。

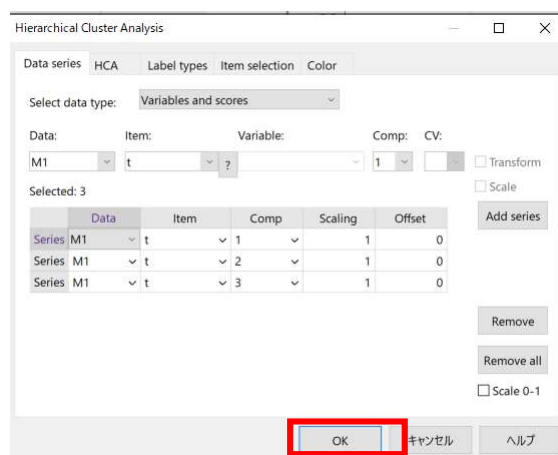
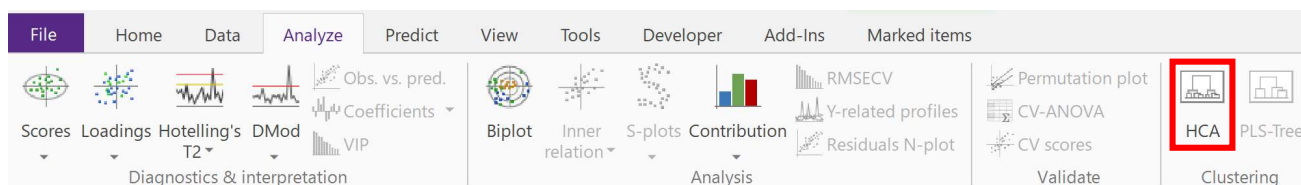


クリスマスブレッドと冷凍魚は、北欧（黄色）全体的で高くなっている変数であることがわかり、その中でも特にスウェーデンにおける特徴的な変数であると解釈することが出来ます。

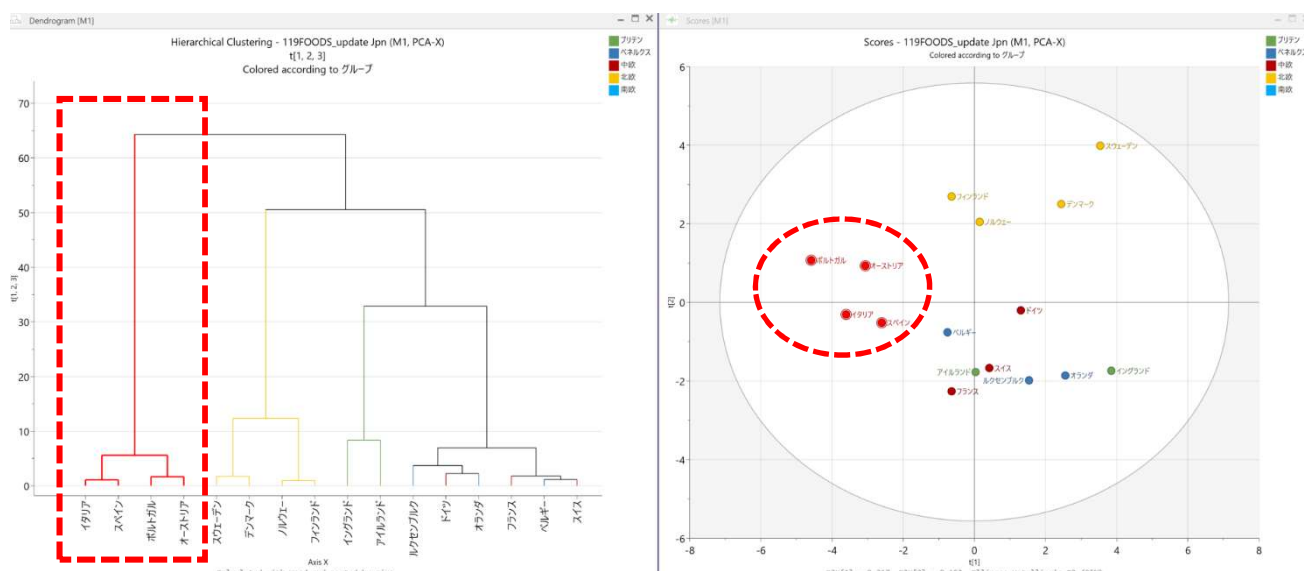


スコアプロットとローディングプロットの互いとスポットの位置関係を比較すると似ていることもわかります。

HCA（階層的クラスターリング）を併用することで、PCA の解釈をサポートすることが出来ます。Analyze > HCA を選択し、Hierarchical Cluster Analysisを開きます。設定はそのまま OK をクリックします。



階層的クラスタリングの結果が表示されます。これとスコアプロットを並べて表示します。階層的クラスタリングで分けられたグループとスコアプロットにおいて距離が近い傾向が見受けられます（例：クラスタリングの左端の群になるイタリア、スペイン、ポルトガル、オーストリアはスコアプロットにおいても距離が近い）



## 3.OPLS

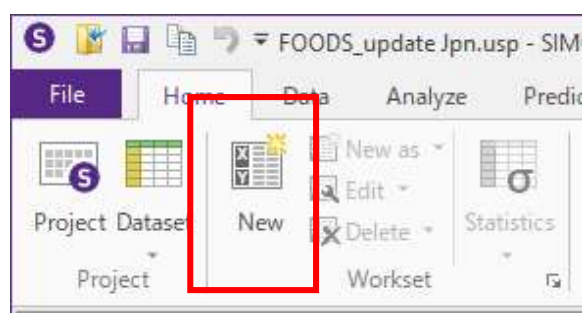
OPLS では Y 変数を設定し、X と Y の回帰式を作成します。

目的変数 Y に連動する、説明変数 X を探すだけではなく、OPLS モデルを使って Y を予測することも可能です。

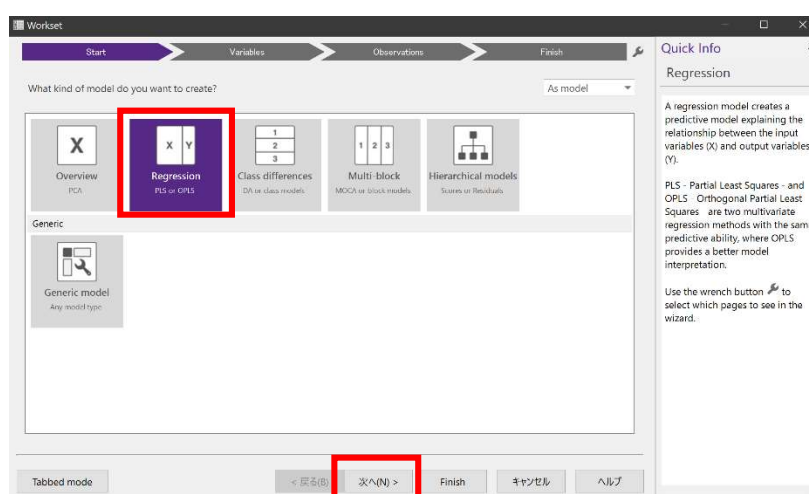
### モデルの作成

#### モデルを設定する

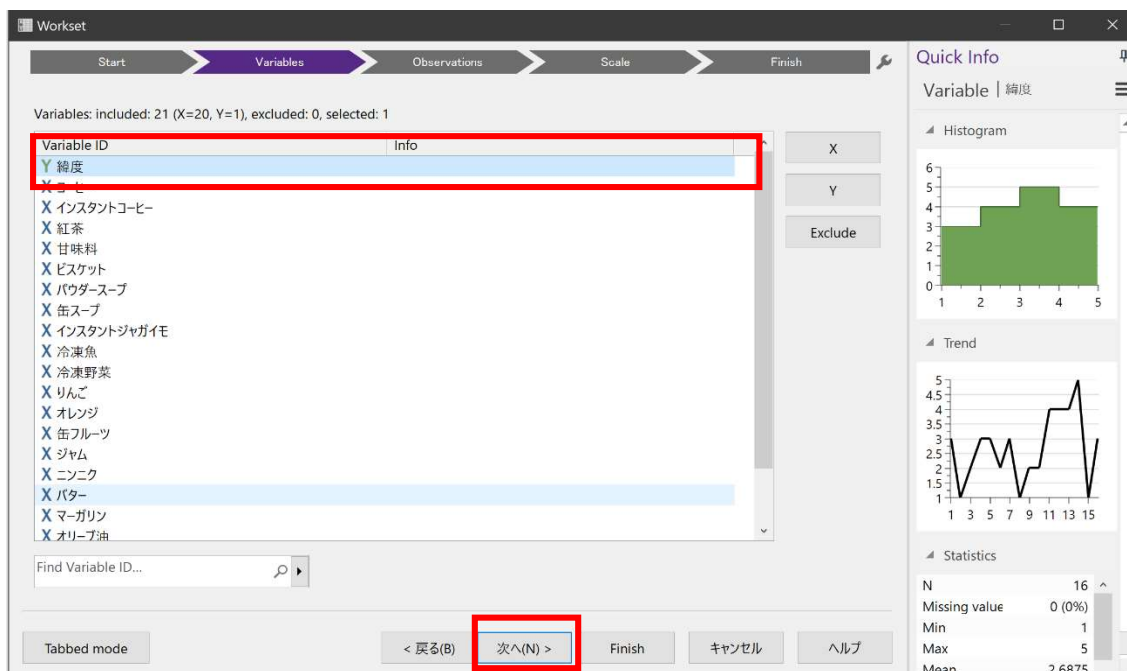
Home > New を選択し、Workset window を開きます。



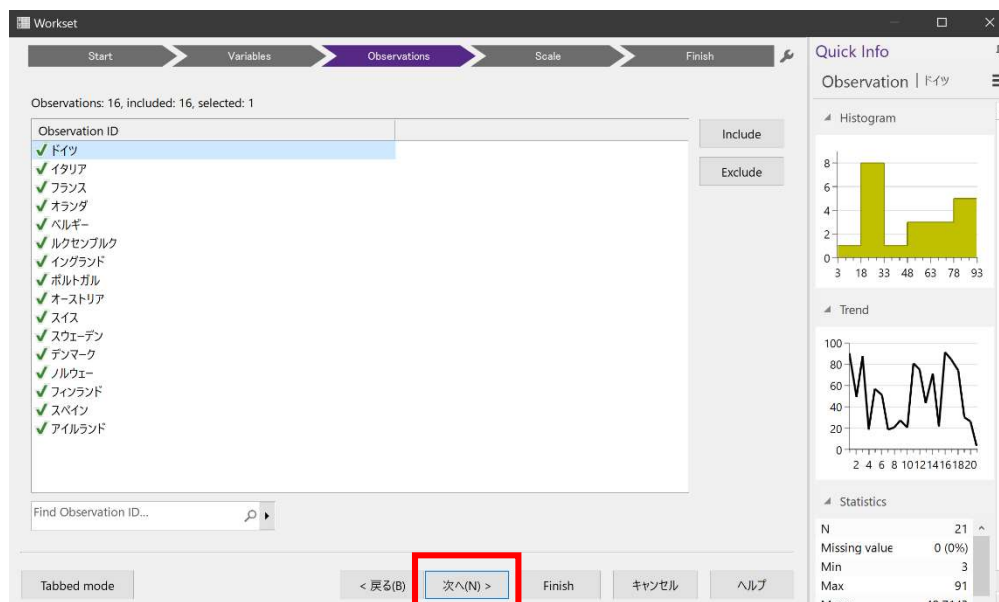
Regression PLS or OPLS を選択し、「次へ」をクリックします。



緯度が Y 変数になっていることを確認し、次へをクリックします。



Observation ID は変更ありません。次へをクリックします。



Scale（スケーリング：尺度を合わせる）の設定をします。不要なサンプルがあれば Type が UV になっていることを確認します（スケーリングは、MS メタボロミクスデータを使った解析編で説明いたします）



Variables: 21, selected: 1

Variable ID	Type	Block	Modifier	Avg.	Std. dev.
* コーヒー	UV	---	1	78.56	23.15
* インスタントコーヒー	UV	---	1	39.25	23.15
* 紅茶	UV	---	1	78.5	18.54
* 甘味料	UV	---	1	18	10.9
* ビスケット	UV	---	1	60.67	19.84
* パウダースープ	UV	---	1	49	15.43
* 缶スープ	UV	---	1	18.31	20.18
* インスタントジャガイモ	UV	---	1	12.75	9.574
* 冷凍魚	UV	---	1	21.88	13.98
* 冷凍野菜	UV	---	1	15.88	12.79
* リンゴ	UV	---	1	66.81	17.58
* オレンジ	UV	---	1	70.5	14.03
* 缶フルーツ	UV	---	1	41.94	23.26
* ジャム	UV	---	1	55.19	22.59
* ニンニク	UV	---	1	42.31	34.68
* バター	UV	---	1	75.81	20.91
* マーガリン	UV	---	1	69.12	26.85
* オリーブ油	UV	---	1	54.19	28.84
* ヨーグルト	UV	---	1	20.53	18.98
* クリスプブレッド	UV	---	1	27.75	25.03
緯度	UV	---	1	2.688	1.195

Set scaling  
Type: UV  
Block: 1  
Block weight: 1/sqrt  
Modifier: 1

Quick Info  
Variable | 緯度 (UV)  
Histogram  
Trend  
Statistics  
N: 16  
Missing values: 0 (0%)  
Min: -1.41157  
Max: 1.93437  
Mean: 1.67638e-08

次へ(N) >

次へをクリックします。

Scale（スケーリング：尺度を合わせる）の設定をします。不要なサンプルがあれば Type が UV になっていることを確認します（スケーリングは、MS メタボロミクスデータを使った解析編で説明いたします）

Variables: 21, selected: 1

Variable ID	Type	Block	Modifier	Avg.	Std. dev.
* コーヒー	UV	---	1	78.56	23.15
* インスタントコーヒー	UV	---	1	39.25	23.15
* 紅茶	UV	---	1	78.5	18.54
* 甘味料	UV	---	1	18	10.9
* ビスケット	UV	---	1	60.67	19.84
* パウダースープ	UV	---	1	49	15.43
* 缶スープ	UV	---	1	18.31	20.18
* インスタントジャガイモ	UV	---	1	12.75	9.574
* 冷凍魚	UV	---	1	21.88	13.98
* 冷凍野菜	UV	---	1	15.88	12.79
* リンゴ	UV	---	1	66.81	17.58
* オレンジ	UV	---	1	70.5	14.03
* 缶フルーツ	UV	---	1	41.94	23.26
* ジャム	UV	---	1	55.19	22.59
* ニンニク	UV	---	1	42.31	34.68
* バター	UV	---	1	75.81	20.91
* マーガリン	UV	---	1	69.12	26.85
* オリーブ油	UV	---	1	54.19	28.84
* ヨーグルト	UV	---	1	20.53	18.98
* クリスプブレッド	UV	---	1	27.75	25.03
緯度	UV	---	1	2.688	1.195

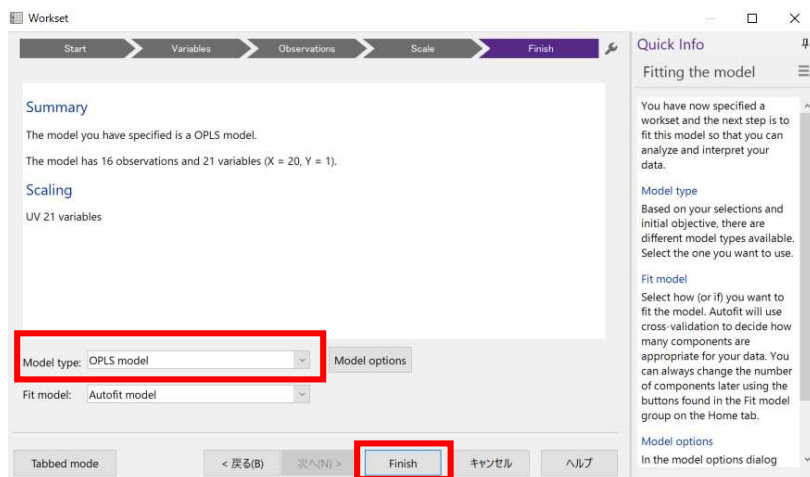
Set scaling  
Type: UV  
Block: 1  
Block weight: 1/sqrt  
Modifier: 1

Quick Info  
Variable | 緯度 (UV)  
Histogram  
Trend  
Statistics  
N: 16  
Missing values: 0 (0%)  
Min: -1.41157  
Max: 1.93437  
Mean: 1.67638e-08

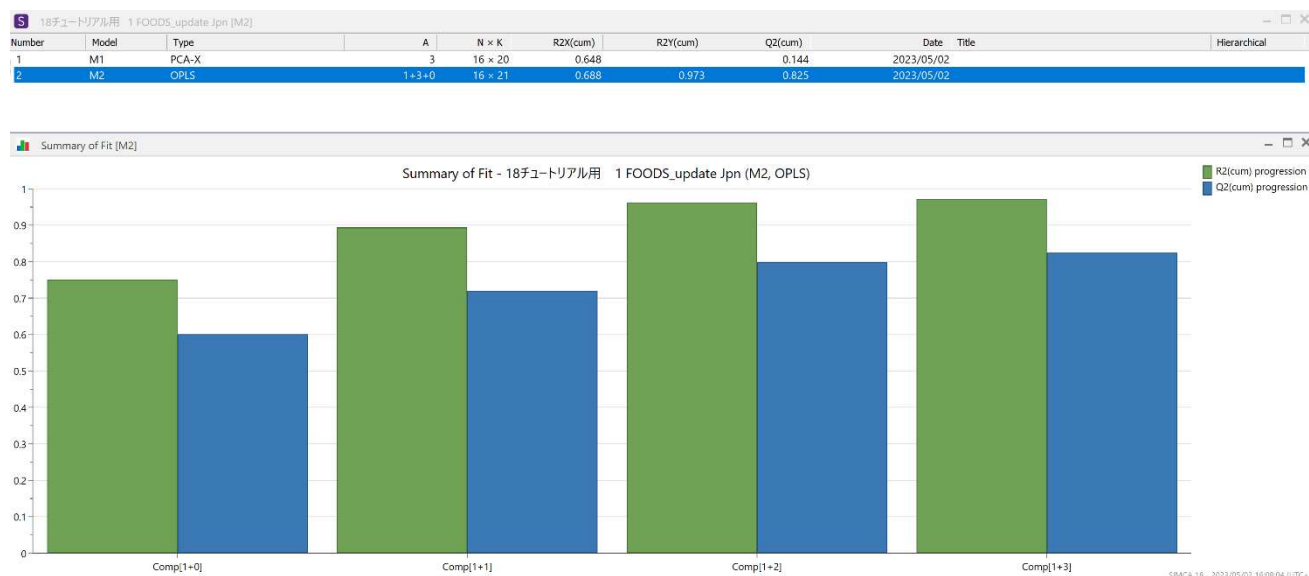
次へ(N) >

次へをクリックします。

Model types を “OPLS model” に設定し、Finish をクリックします。



モデルが作成されます。

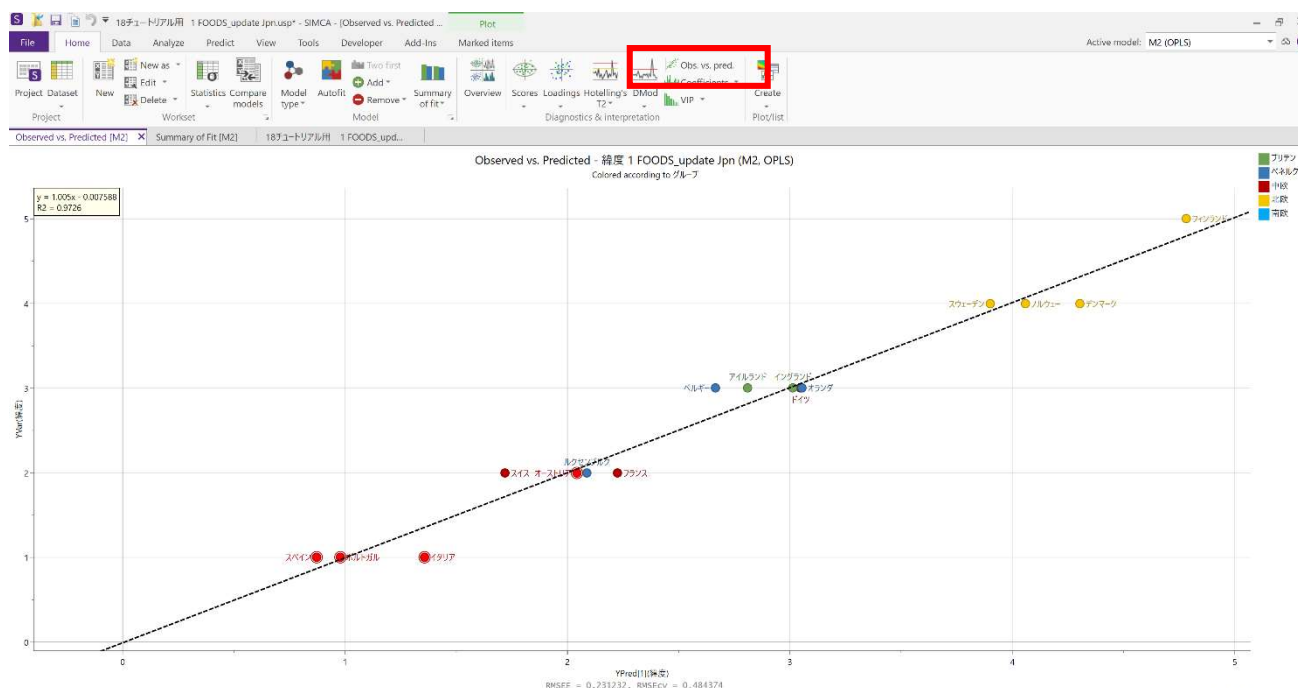


R2（緑）は第 4 成分までのモデルで説明できた Y の変動を表し、Q2（青）は交差検証に従ってこのモデルで予測出来た Y の変動を表します。

## OPLS 回帰モデル

Y 変数（ここでは緯度）を予測する、回帰モデルを作成されました。予測と実測の比較し、精度を確認します。

Home > Obs. vs. pred. をクリックします。

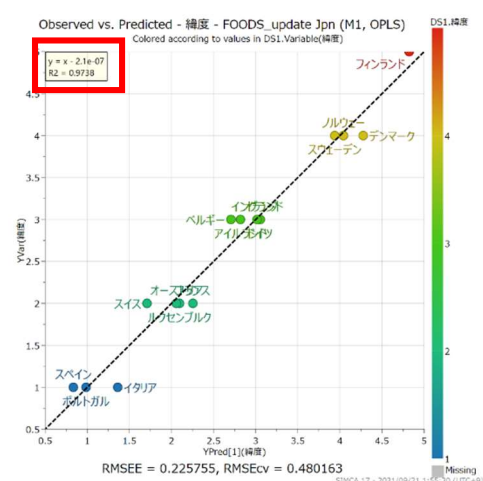


YPred は作成したモデルの回帰式に当てはめた、理論上の Y の予測値、Yvar は Y の実測値を示します。破線で示される回帰直線は残差(観測値と予測値との差)の二乗和が最小となる最小二乗法によって求められます。

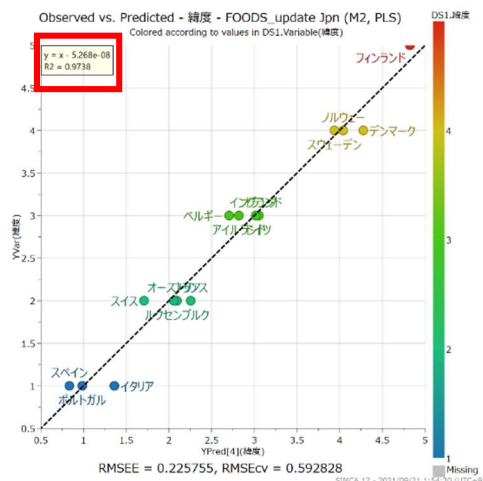
この予測の良さが、先程 Comp[1+3]で表されていた  $R^2=0.9726$  であり、非常に良い予測モデルと言えます。

💡 OPLS と PLS の予測値や決定係数 ( $R^2$ ) は、データセットに欠損値がある場合を除いて同じになります。

### OPLS



### PLS



Copyright Infocom Corporation 本資料の無断複製および転用は禁じられています



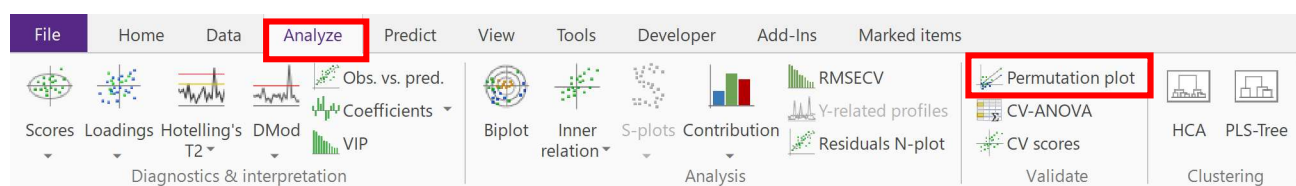
## モデルの診断

並べ替え検定（permutation test）を実施し、モデルが過剰適合しているか否かを診断します。

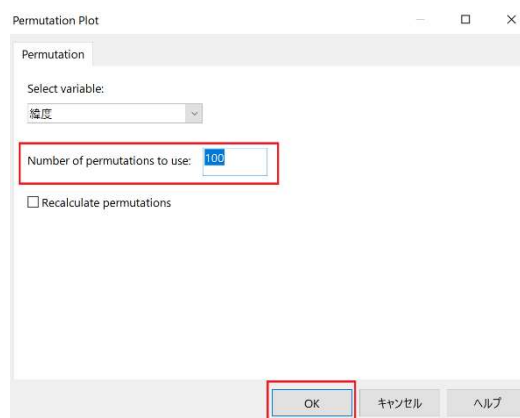
過剰適合（overfitting）とは？

モデルが複雑過ぎることで発生し、関係のない特徴にまで適合してしまう状態です。

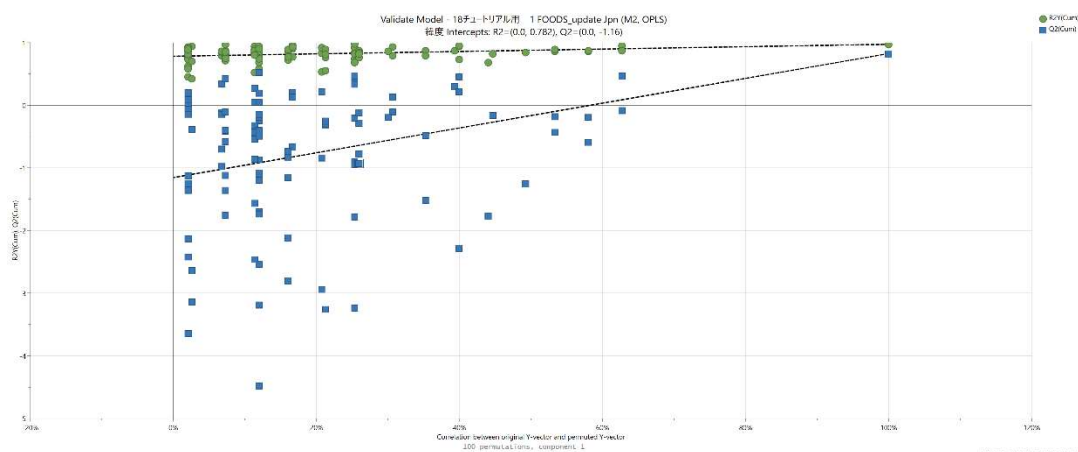
Analyze > Permutation plot をクリックします。



Permutation plot の設定が表示されます。Number of permutations to use を 100 に設定し、OK をクリックします（Number of permutations to use は、Y の値をサンプル間で並べ替え計算をする回数を設定します。）

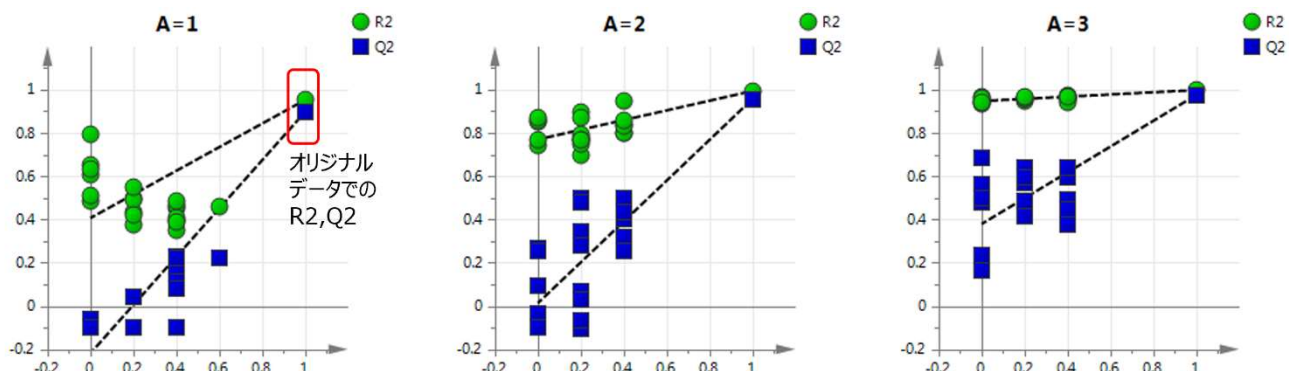


以下の結果が表示されます。



💡 各(R2,Q2)の組が Y を入れ替えたそれぞれのモデルに対応しています。横軸は Y と Y' の相関係数を示しています。

成分数 A を増加し、モデルが複雑化すると、R2、Q2 の傾きが緩くなります。このような状態は、どんな壊れたデータでも良いモデルが得られる過剰適合となります。



Q2 の Y 切片がゼロ以下、R2 も適度に左下がりになっているものが適当なモデルの目安としてお考え下さい。

## 結果の解釈

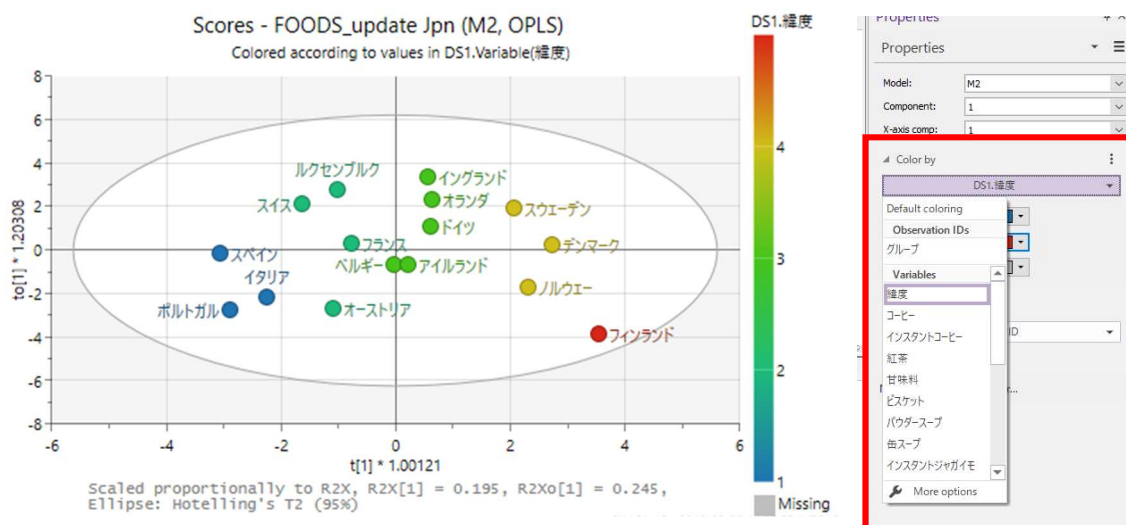
### スコアプロット

Home > Score をクリックします。

Y 変数が一つの場合の OPLS では、横軸は Y の変動を表し、縦軸は Y が同じグループ内の変動を表します。



右クリック>Properties > Color by を緯度に変更します。



左から右に緯度が大きくなる変化に従って、国々が分類されているのが確認できます（色が青から赤）。

Copyright Infocom Corporation 本資料の無断複製および転用は禁じられています

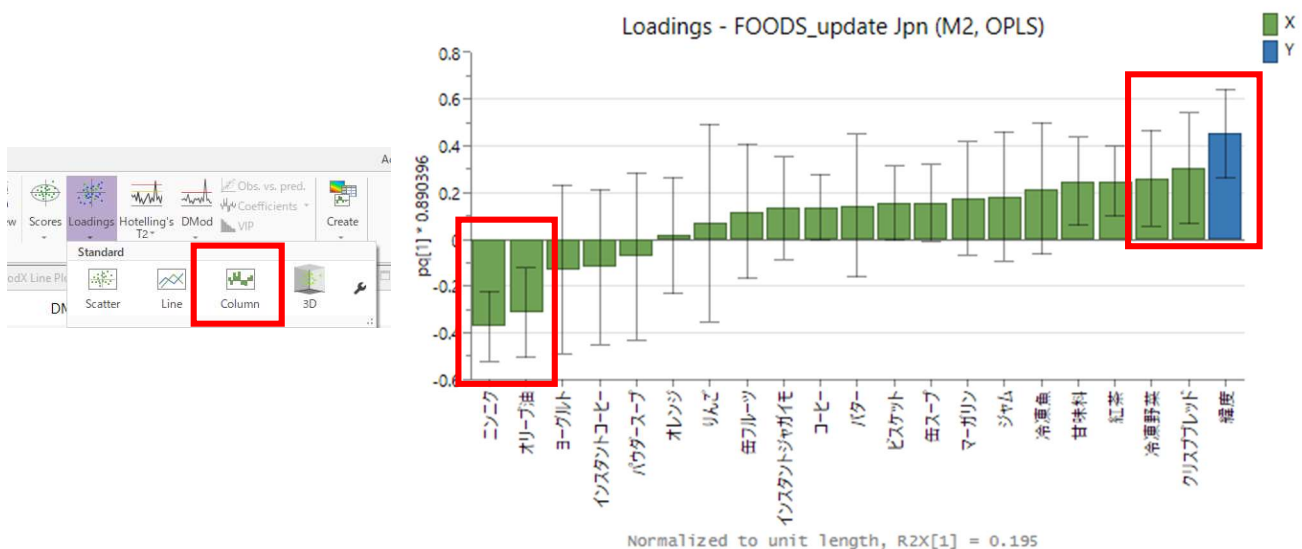


## ローディングプロット

Home > Loadings をクリックします。先程と同様に、ローディングプロットはスコアプロットとの相関を表します。

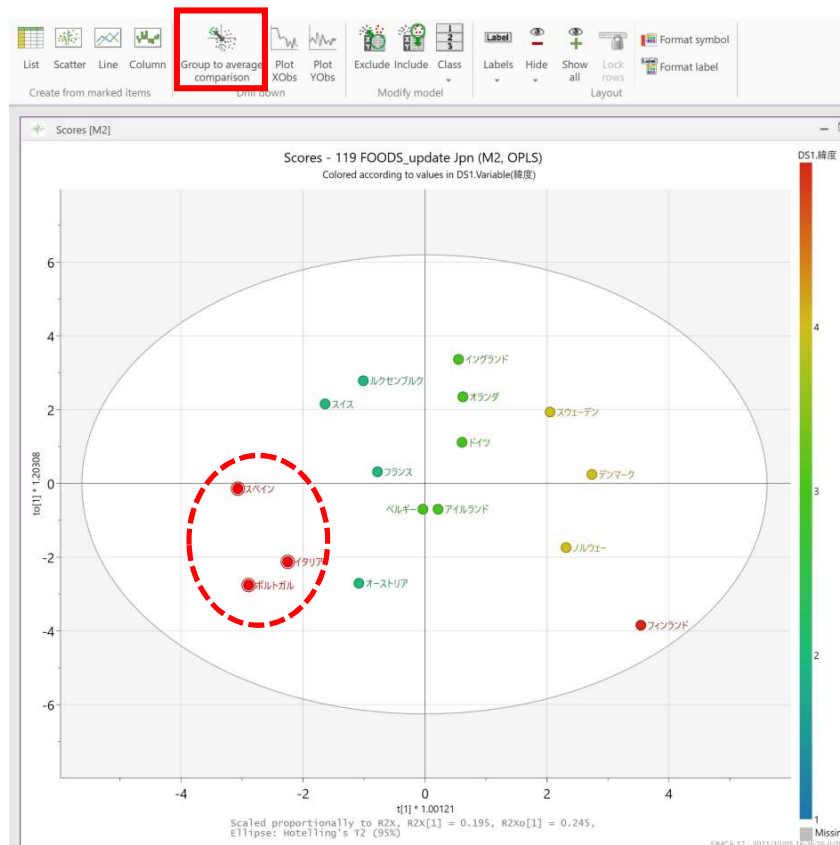


Home > Loadings > Column をクリックします。表示されたグラフを、PCA の時と同様に、右クリック→Sort ascenbling で並び替えを行ってください。以下のようなグラフが作成されます。



緯度はクロシブレッドや冷凍野菜と正の相関が高く、ニンニクやオリーブオイルとは負の相関が高いことがわかります。つまり、緯度が高い国では、クロシブレッドや冷凍野菜の消費量が多く、ニンニクやオリーブオイルの消費量が低い関係を持っているということが、この結果から解釈することができます。

スコアプロット上で緯度の低いスペイン、ポルトガル、イタリアを選択し Marked items > Group to average comparison をクリックします。



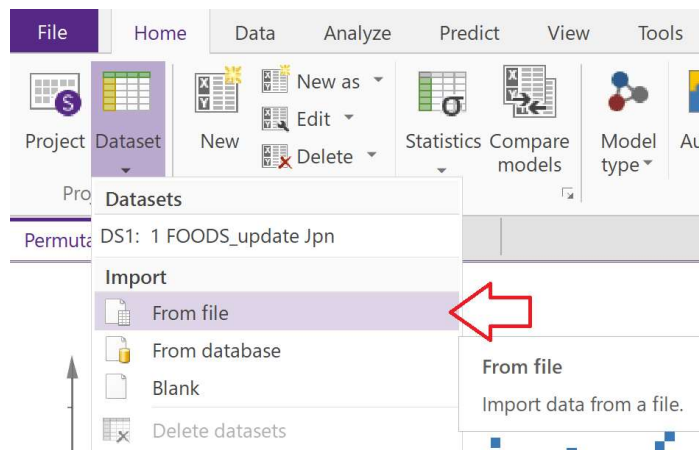
緯度が低いほどオリーブ油とにんにくの消費は増加しますが、3 国の消費の平均から全体の平均を引くとその値が正になっています。



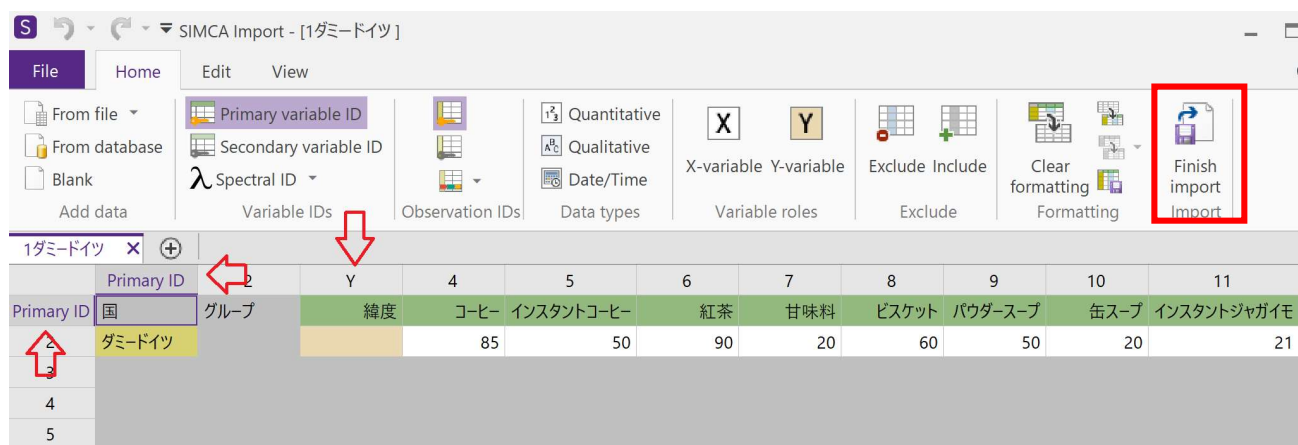
## 回帰モデルを用いた予測

### データの読み込みと予測

Home>Dataset>Import: From file から 1 ダミードイツ .xlsx を読み込みます。



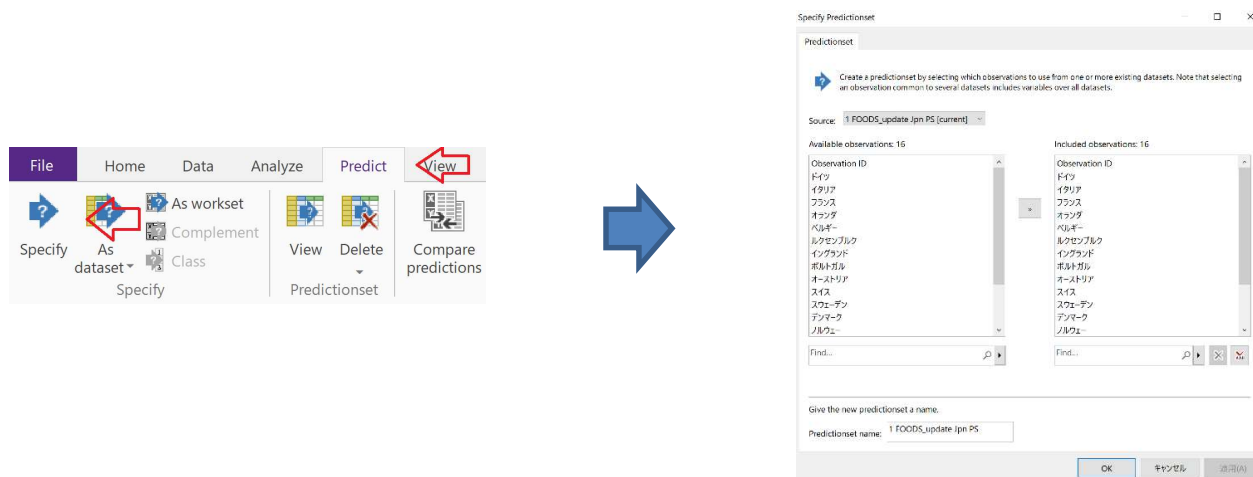
データを下記のように設定し（矢印参照）、完了後に Finish import をクリックします



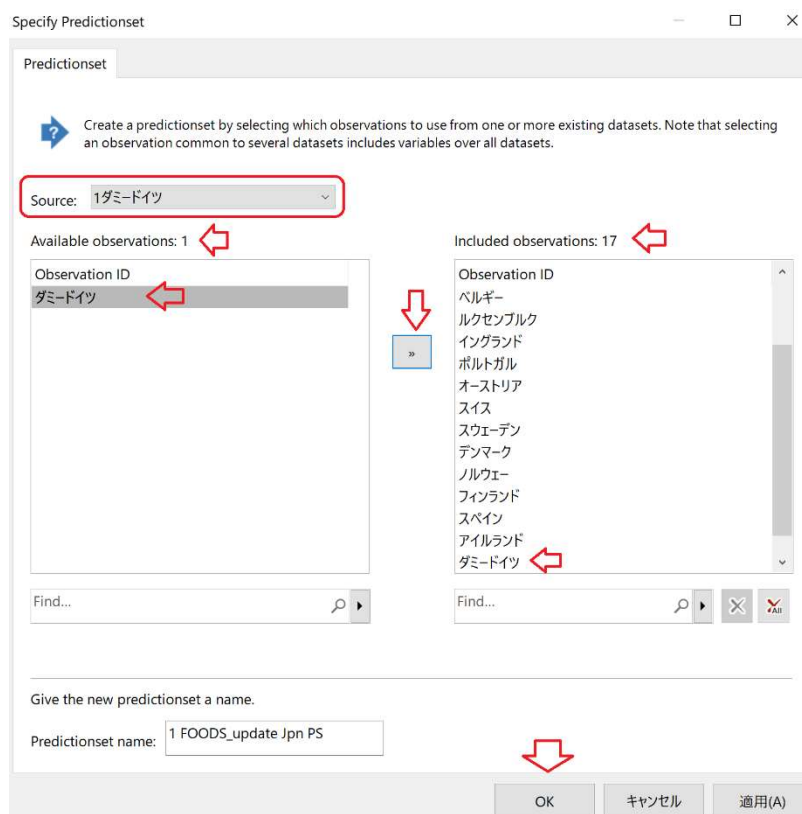
読み込み後のデータが下記のように表示されます（Y 値が空欄です）

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Primary ID	緯度	コーヒー	インスタントコーヒー	紅茶	甘味料	ビスケット	パウダースープ	缶スープ	インスタントジャガイモ	冷凍魚	冷凍野菜	りんご	オレンジ
2	ダミードイツ		85	50	90	20	60	50	20	21	27	21	81	75

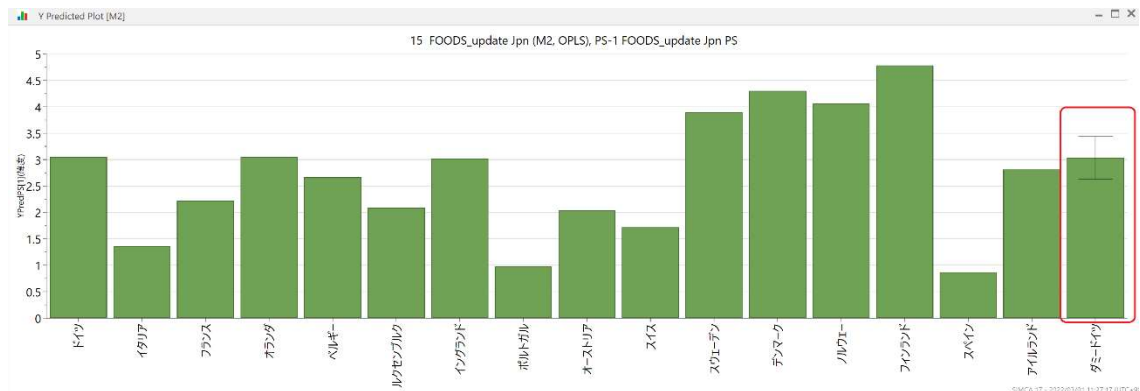
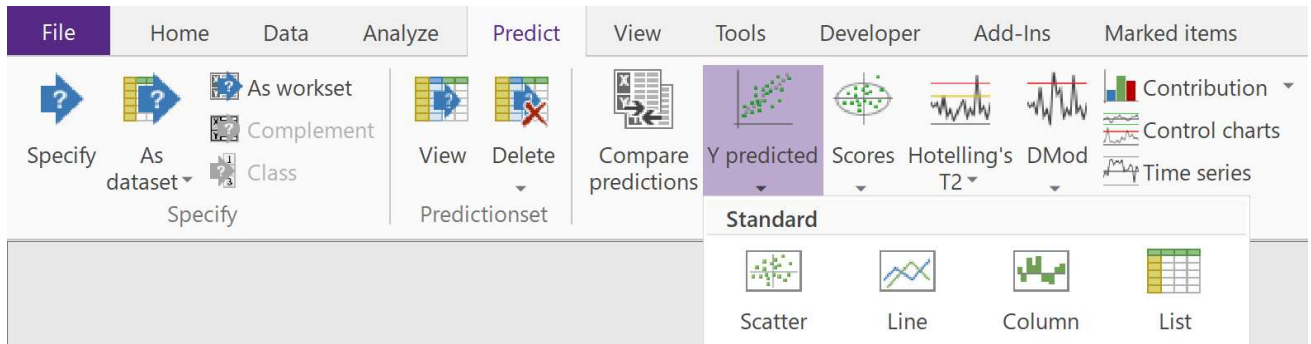
Predict>Specify をクリックし、Specify Predictionset ウィンドウを開きます。



Source で 1 ダミードイツを選択し Available observations でダミードイツを選択します。『>』アイコンをクリックし、include observations に追加します。追加後に OK ボタンをクリックします。



Predict>Y predicted>Column でダミードイツの予測値が表示されます。



Sort ascending で並べ替えるとドイツの実測値とダミードイツの予測値が近いことが分かり易いです。

