

405. データマイニング実験

1. 目的

データマイニングとは、データをマイニング（発掘）して、情報・知見・知識・仮説・課題などを見つける手法・プロセスのことである。いくつかの課題を通じて実際のデータを解析することによって、データマイニングの中心的な手法である回帰分析について理解する。

2. 例題

表1に示したデータは、それぞれの胸囲、身長、体重の関係を示している。ここでは、胸囲、および身長が、体重に対してどのような関係があるかを分析する。

表1：胸囲、身長、体重のデータ

No.	胸囲(cm)	身長(cm)	体重(kg)
1	80	164	53
2	82	158	55
3	85	162	56
4	82	158	51
5	84	164	56
6	82	152	44
7	77	153	50
8	78	157	47
9	78	153	45
10	82	158	54
11	75	160	51
12	82	150	51

分析手順は、以下のようである。

[手順1]：Excel上に表を作成する

[手順2]：それぞれの変数に対して散布図を作成する

[手順3]：単回帰式、および、相関係数を求める。

[手順4]：単回帰式を利用して残差と相対誤差を求める。

[手順5]：重回帰分析を行う。

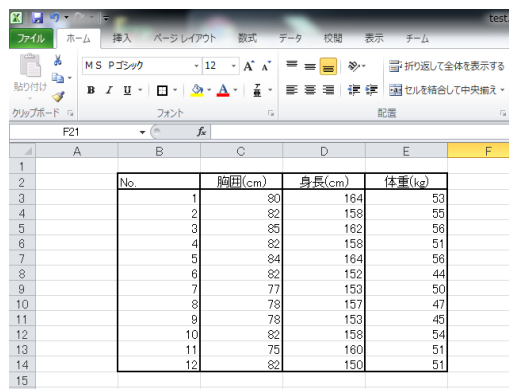
[手順6]：重回帰式、および重相関係数を求める。

[手順7]：重回帰式を利用して残差と相対誤差を求める。

以下では、それぞれの手順の詳細について説明する。

2. 1. [手順 1]: Excel 上に表の作成

Excel を起動し、表 1 と全く同じように、Excel 上に表を作成する (図 1 参照)。



No.	胸囲(cm)	身長(cm)	体重(kg)
1	80	164	53
2	82	158	55
3	85	162	56
4	82	158	51
5	84	164	56
6	82	152	44
7	77	153	50
8	78	157	47
9	78	153	45
10	82	158	54
11	75	180	51
12	82	150	51

図 1 : 表の作成

2. 2. [手順 2]: それぞれの変数に対して散布図の作成

まず、胸囲と体重の関係についての散布図を作成する。

Excel の「挿入タブ」を選択し、「散布図」の中の左上の散布図を選択する (図 2 参照)。

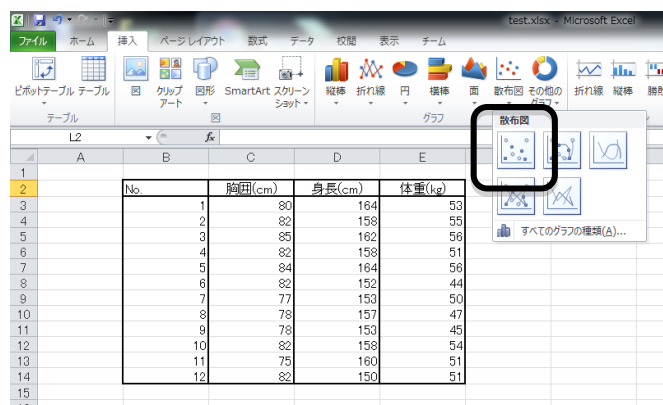


図 2 : 散布図の選択

図 3 のようなグラフを表示するための白いウィンドウが表示される。グラフツールの「デザインタブ」の「データの選択」をクリックする (※グラフのウィンドウを選択しないとグラフツールのタブは表示されない)。

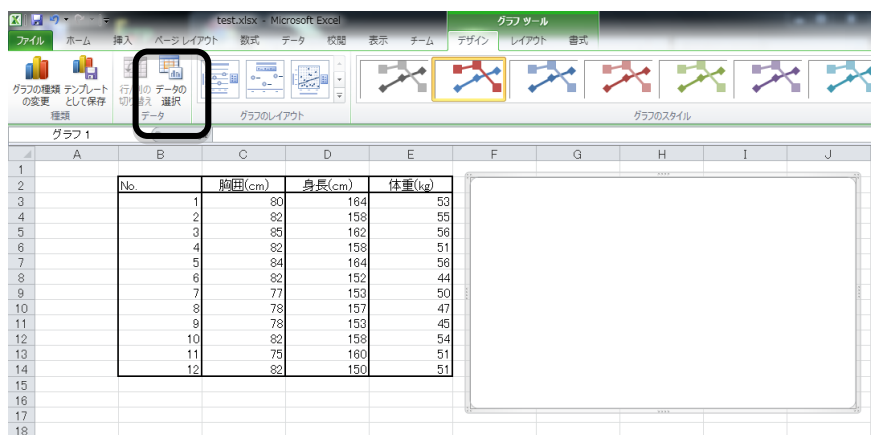


図 3 : データの選択

図 4 のようなデータソースの選択ウィンドウが表示される。「追加ボタン」をクリックする。

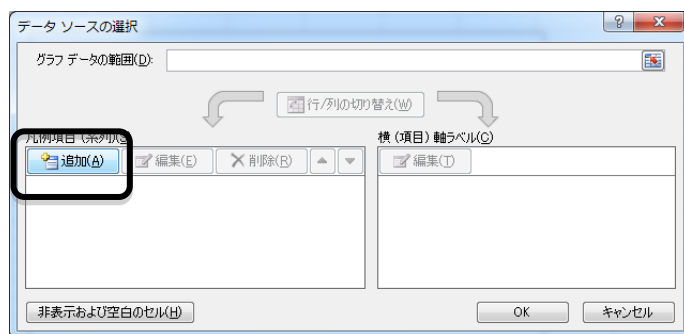


図 4 : データソースの選択

図 5 のような系列の編集ウィンドウが表示される。「系列名」(胸囲(cm)), 「系列 X の値」(胸囲のデータ), 「系列 Y の値」(体重のデータ) を, それぞれのデータセルを選択することで指定し, 「OK ボタン」を押す。

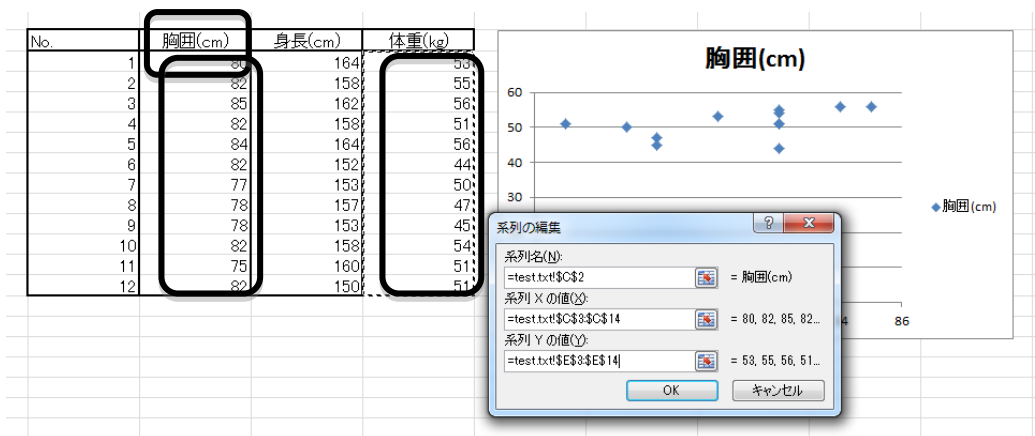


図 5 : 系列の編集

「OK ボタン」を押すと, 図 6 左のような胸囲と体重の関係を示した散布図が表示される。同様の手順で, 身長と体重の関係を示した散布図を作成すると, 図 6 右のような散布図が表示される。

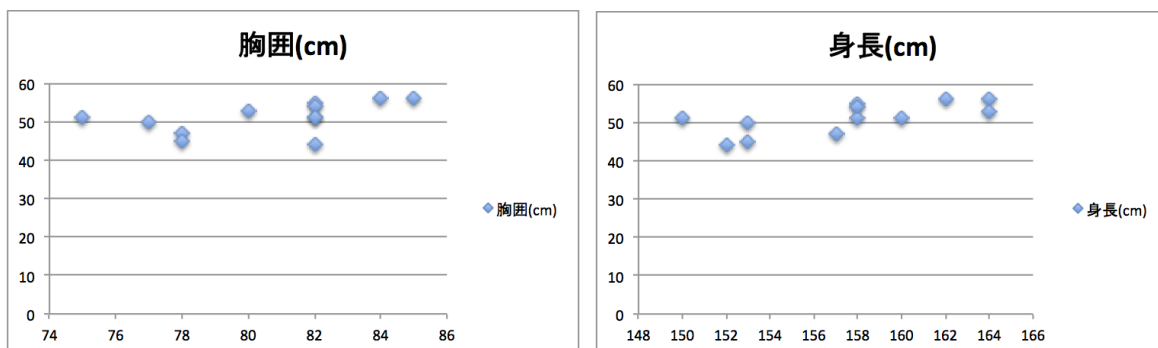


図 6 : 胸囲と体重の関係 (左) と身長と体重の関係 (右)

これらの散布図を見ると, 胸囲および身長に従って体重が多くなるような, 比例の関係を確認することができる。また, 若干ではあるが, 胸囲と身長では, 身長の方が, より比例的な関係にあるように見受けられる。

2. 3. [手順3]: 単回帰式, および相関係数の導出

グラフツールの「レイアウトタブ」の「近似曲線」を選択し, さらに「その他の近似曲線オプション」を選択する.

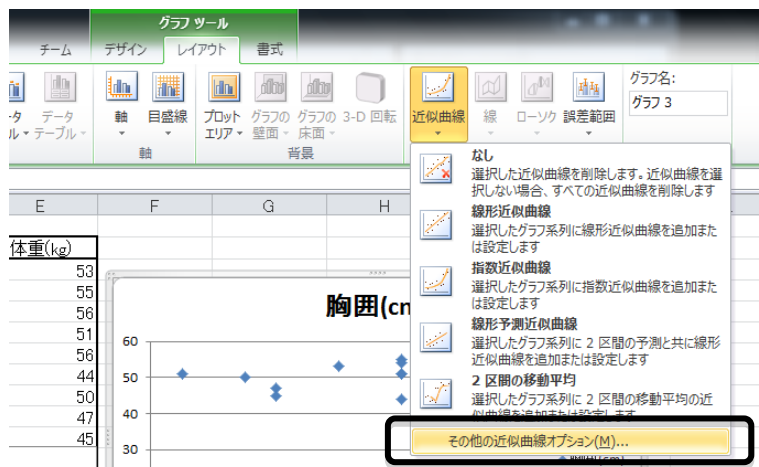


図7: 近似曲線の選択

すると, 図8のような近似曲線の書式設定ウィンドウが表示されるので, 「線形近似」, 「グラフに数式を表示する」, 「グラフに R-2 乗値を表示する」を選択して, 「閉じるボタン」を押す.

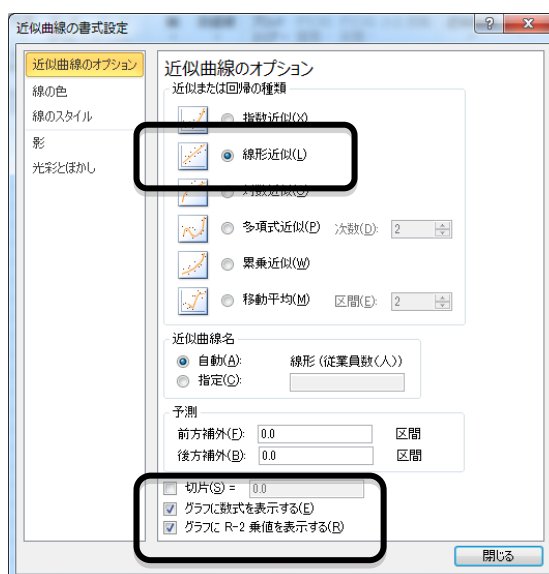


図8: 近似曲線の書式設定

図9および図10のように, グラフ上に表示された数式が近似直線の式であり, 胸囲 x と体重 y との関係, および, 身長 x と体重 y との関係を示した式である. つまり, 胸囲や身長から体重を以下のように計算によって推定することができる.

$$\text{体重 (kg)} = 0.6816 \times \text{胸囲 (cm)} - 3.8382 \quad (1)$$

$$\text{体重 (kg)} = 0.601 \times \text{身長 (cm)} - 43.52 \quad (2)$$

これらの式を「回帰式」といい, 特に変数がそれぞれ1つなので「単回帰式」という. ちなみに, 手順5以降の重回帰分析による回帰式は, 変数が複数なので「重回帰式」という.

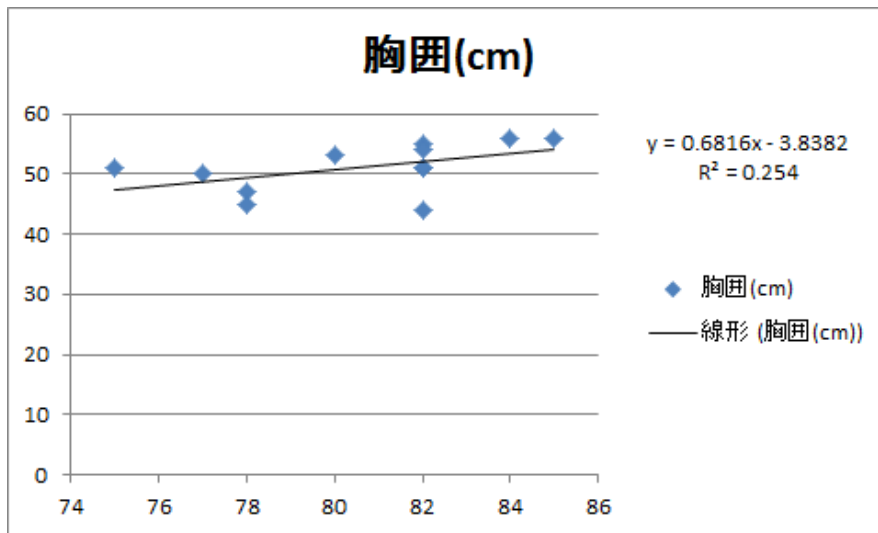


図 9 : 胸囲と体重における近似直線, 回帰式, 相関係数の表示

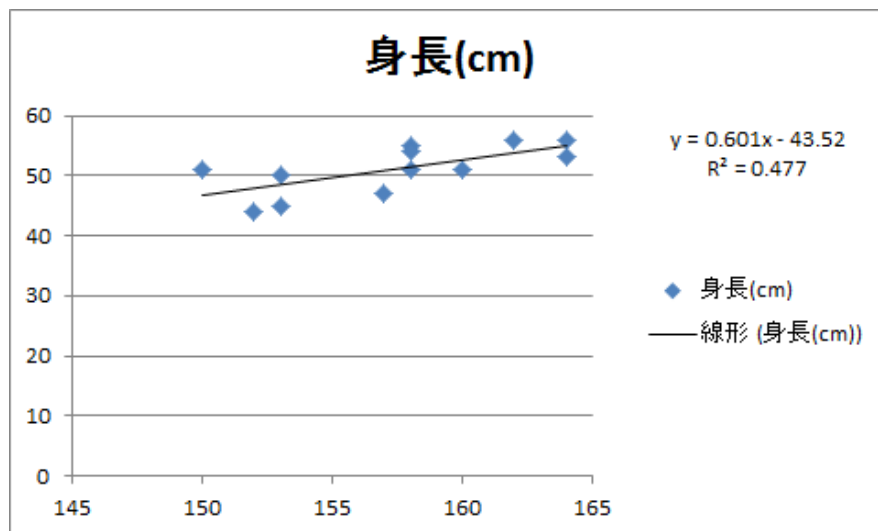


図 10 : 胸囲と体重における近似直線, 回帰式, 相関係数の表示

また, グラフ上に表示されている R^2 は, 相関係数を 2 乗した値を示している. 相関係数とは, それぞれの値にどの程度の関係性があるかを示す指標であり, -1 から 1 の範囲の実数値である. 1 に近いときは正の相関 (一方が増加すれば他方も増加する関係) があるといい, -1 に近ければ負の相関 (一方が増加すれば他方は減少する関係) があるという. 0 に近いときは相関が弱いことを意味する. したがって, R^2 の値がより大きな値の時は, それぞれの関係性が強いことを意味する.

胸囲の場合の R^2 は 0.254 であり, 身長の場合の R^2 の値は 0.477 である. つまり, 身長の方が体重に対して関係性が強いことを意味している. これは手順 2 の段階の散布図からも見受けられたが, それを実際の数値として表したものである.

2. 4. [手順 4]: 単回帰式を利用した残差と相対誤差の計算

手順 3 で求めた単回帰式 (1) および (2) に対して, 表 1 のそれぞれの値を代入して体重の推定値を求める. そして, 実際の値 (観測値) との差 (残差) を求め, それぞれの相対誤差を計算する. ここで, 「残差 = 観測値 - 推定値」であり, 「相対誤差 = 残差 / 観測値 $\times 100$ 」で計算する. 最後に, 相対誤差の絶対値の平均を求め, それぞれの変数に対して, 表 2 と表 3 のような表を作成する.

これらの結果からも, 身長による指定値の方が, 相対誤差が小さく精度が高いことが分かる.

表 2：胸囲による推定値の残差と相対誤差

No.	観測値: 体重(kg)	推定値 体重(kg)	残差	相対誤差(%)	相対誤差 の絶対値(%)
1	53	50.6898	2.3102	4.358867925	4.358867925
2	55	52.053	2.947	5.358181818	5.358181818
3	56	54.0978	1.9022	3.396785714	3.396785714
4	51	52.053	-1.053	-2.064705882	2.064705882
5	56	53.4162	2.5838	4.613928571	4.613928571
6	44	52.053	-8.053	-18.30227273	18.30227273
7	50	48.645	1.355	2.71	2.71
8	47	49.3266	-2.3266	-4.950212766	4.950212766
9	45	49.3266	-4.3266	-9.614666667	9.614666667
10	54	52.053	1.947	3.605555556	3.605555556
11	51	47.2818	3.7182	7.290588235	7.290588235
12	51	52.053	-1.053	-2.064705882	2.064705882
				平均	5.694205979

表 3：身長による推定値の残差と相対誤差

No.	観測値: 体重(kg)	推定値 体重(kg)	残差	相対誤差(%)	相対誤差 の絶対値(%)
1	53	55.044	-2.044	-3.856603774	3.856603774
2	55	51.438	3.562	6.476363636	6.476363636
3	56	53.842	2.158	3.853571429	3.853571429
4	51	51.438	-0.438	-0.858823529	0.858823529
5	56	55.044	0.956	1.707142857	1.707142857
6	44	47.832	-3.832	-8.709090909	8.709090909
7	50	48.433	1.567	3.134	3.134
8	47	50.837	-3.837	-8.163829787	8.163829787
9	45	48.433	-3.433	-7.628888889	7.628888889
10	54	51.438	2.562	4.744444444	4.744444444
11	51	52.64	-1.64	-3.215686275	3.215686275
12	51	46.63	4.37	8.568627451	8.568627451
				平均	5.076422748

表 2 と表 3 を作成するに当たって、エクセルにおける計算の方法を簡単に説明しておく。エクセルにおいて、最初に「=」を記述することで、計算式を指定することができる。図 1 1 に示したように、「=」の後に、実際の計算式を指定することで、その計算を行った結果の値をそのセルに表示させることができる。これは、先ほど求めた単回帰式 (1) を意味しており、変数 x に当たる部分には、実際に代入する値の書かれたセルの番号（ここでは B20）を指定することが可能である。セルの番号を入力する際には、そのセルをクリックすることで番号が自動的に入力される。

一度作成した計算式をコピー&ペーストすることも可能である。その場合、セルの番号が自動的にずれてコピーされるため、一つ一つ入力する必要はなく、非常に便利である。

No.	胸囲(cm)	身長(cm)	体重(kg)	No.	観測値: 体重(kg)	推定値 体重(kg)	残差	相対誤差(%)	相対誤差 の絶対値(%)
1	80	164	53	1	53	=0.68165*B20+3.8382		-10.13245283	10.13245283
2	82	158	55	2	55	59.7335	-4.7335	-8.606363636	8.606363636
3	85	162	56	3	56	61.77845	-5.77845	-10.31866071	10.31866071
4	82	158	51	4	51	59.7335	-8.7335	-17.1245098	17.1245098
5	84	164	56	5	56	61.0968	-5.0968	-9.101428571	9.101428571
6	82	152	44	6	44	59.7335	-15.7335	-35.75795455	35.75795455
7	77	153	50	7	50	56.32525	-6.32525	-12.6505	12.6505
8	78	157	47	8	47	57.0069	-10.0069	-21.2912766	21.2912766
9	78	153	45	9	45	57.0069	-12.0069	-26.682	26.682
10	82	158	54	10	54	59.7335	-5.7335	-10.61759259	10.61759259
11	75	160	51	11	51	54.96195	-3.96195	-7.768529412	7.768529412
12	82	150	51	12	51	59.7335	-8.7335	-17.1245098	17.1245098
平均				平均					15.59798154

図 1 1 : エクセルにおける計算式の入力

No.	観測値: 体重(kg)	推定値 体重(kg)	残差	相対誤差(%)	相対誤差 の絶対値(%)
1	53	58.3702	-5.3702	-10.13245283	=ABS(J20)
2	55	59.7335	-4.7335	-8.606363636	8.606363636
3	56	61.77845	-5.77845	-10.31866071	10.31866071
4	51	59.7335	-8.7335	-17.1245098	17.1245098
5	56	61.0968	-5.0968	-9.101428571	9.101428571
6	44	59.7335	-15.7335	-35.75795455	35.75795455
7	50	56.32525	-6.32525	-12.6505	12.6505
8	47	57.0069	-10.0069	-21.2912766	21.2912766
9	45	57.0069	-12.0069	-26.682	26.682
10	54	59.7335	-5.7335	-10.61759259	10.61759259
11	51	54.96195	-3.96195	-7.768529412	7.768529412
12	51	59.7335	-8.7335	-17.1245098	17.1245098
平均					15.59798154

No.	観測値: 体重(kg)	推定値 体重(kg)	残差	相対誤差(%)	相対誤差 の絶対値(%)
1	53	58.3702	-5.3702	-10.13245283	10.13245283
2	55	59.7335	-4.7335	-8.606363636	8.606363636
3	56	61.77845	-5.77845	-10.31866071	10.31866071
4	51	59.7335	-8.7335	-17.1245098	17.1245098
5	56	61.0968	-5.0968	-9.101428571	9.101428571
6	44	59.7335	-15.7335	-35.75795455	35.75795455
7	50	56.32525	-6.32525	-12.6505	12.6505
8	47	57.0069	-10.0069	-21.2912766	21.2912766
9	45	57.0069	-12.0069	-26.682	26.682
10	54	59.7335	-5.7335	-10.61759259	10.61759259
11	51	54.96195	-3.96195	-7.768529412	7.768529412
12	51	59.7335	-8.7335	-17.1245098	17.1245098
平均					=AVERAGE(K20:K31)

図 1 2 : エクセルにおける関数の入力例 (左 : 絶対値 ABS, 右 : 平均 AVERAGE)

また、エクセルに予め用意された関数を利用することも可能である。図 1 2 は、絶対値を求める関数「ABS」と平均値を求める関数「AVERAGE」を利用した例である。平均において「K20:K30」のように平均したいセルの範囲を指定しているが、これもマウスで選択することで指定することが可能である。表示させるセルにおいて「=AVERAGE(」まで入力し、その後マウスで選択すると自動的にセルの値が入力される。

2. 5. [手順 5]: 重回帰分析

「データタブ」の「データ分析」をクリックすると図 1 3 のようなデータ分析ウィンドウが表示される。そこで、「回帰分析」を選択して「OK ボタン」を押す。

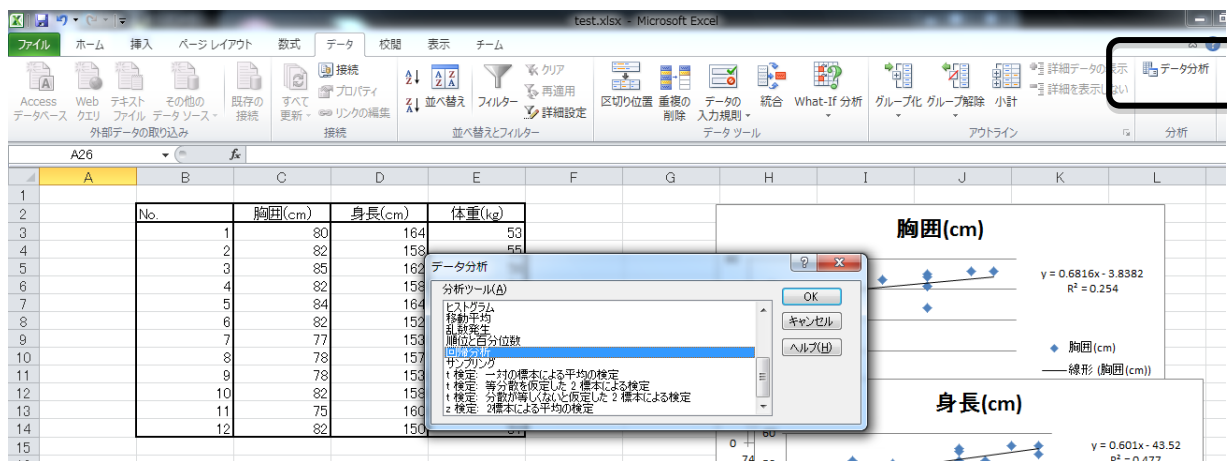


図 1 3 : 回帰分析の選択

次に、図14の「回帰分析ウィンドウ」において、「入力 Y 範囲」でデータの体重の列を選択し、「入力 X 範囲」では、胸囲と身長を選択する。選択する際にはラベルの部分も選択する。また、「ラベル」のチェックも入れておく。

「OK ボタン」を押すと、図15のような分析結果のシートが表示される。

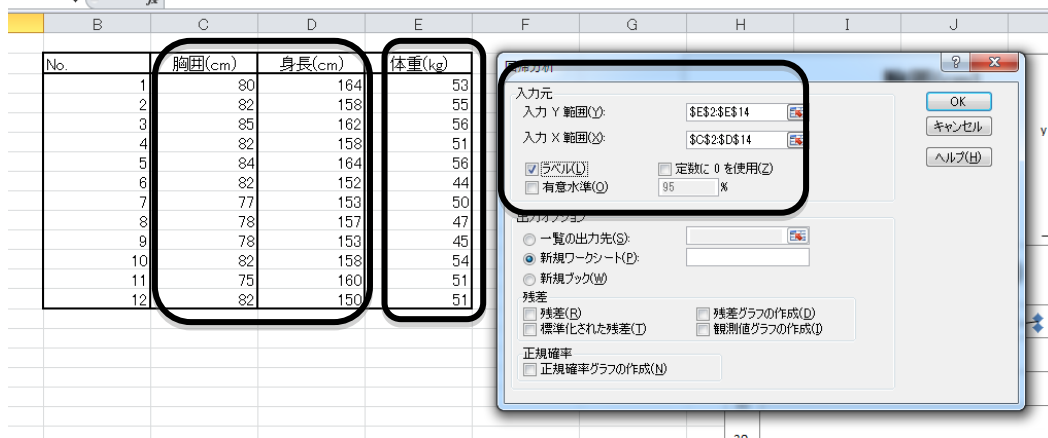


図14：回帰分析の設定

回帰統計	
重相関 R	0.773179
重決定 R ²	0.597608
補正 R ²	0.508429
標準誤差	2.843386
観測数	12

分散分析表						
	自由度	変動	分散	ばれた分散	有意 F	
回帰	2	108.153	54.07652	6.688627	0.016594	
残差	9	72.76362	8.084847			
合計	11	180.9167				

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	-70.7215	33.35915	-2.12	0.063037	-146.185	4.742169	-146.185	4.742169
胸囲(cm)	0.484436	0.294591	1.644437	0.134497	-0.18197	1.150847	-0.18197	1.150847
身長(cm)	0.525785	0.189553	2.773815	0.021618	0.096986	0.954584	0.096986	0.954584

図15：回帰分析の出力

2. 6. [手順6]：重回帰式，および重相関係数の導出

図15の回帰分析の出力の中の「係数」の部分に着目する。これらは回帰式の係数を表しており，以下のような関係を示している。

$$\text{体重(kg)} = 0.484436 \times \text{胸囲(cm)} + 0.52575 \times \text{身長(cm)} - 70.7215 \quad (3)$$

これは，変数の数が1つ以上なので，式(1)や(2)の「単回帰式」に対して「重回帰式」という。また，図15の「重相関 R」の部分为重相関係数であり，重回帰分析における相関係数である。範囲は，0から1の範囲であり，値が大きいほど関係性が強いことを意味する。

2. 7. [手順7]: 重回帰式を利用して残差と相対誤差の計算

表2, 表3と同様に, 手順6で求めた重回帰式にそれぞれの値を代入して体重の推定値を求める. そして, 実際の値(観測値)との差(残差)を求め, それぞれの相対誤差を計算する. ここで, 「残差=観測値-推定値」であり, 「相対誤差=残差/観測値×100」で計算する. 最後に, 相対誤差の絶対値の平均を求め, 表4のような表を作成する.

表4: 重回帰式による推定値の残差と相対誤差

No.	観測値: 体重(kg)	推定値 体重(kg)	残差	相対誤差(%)	相対誤差 の絶対値(%)
1	53	54.26212	-1.26212	-2.381358491	2.381358491
2	55	52.076282	2.923718	5.315850909	5.315850909
3	56	55.63273	0.36727	0.655839286	0.655839286
4	51	52.076282	-1.076282	-2.110356863	2.110356863
5	56	56.199864	-0.199864	-0.3569	0.3569
6	44	48.921572	-4.921572	-11.18539091	11.18539091
7	50	47.025177	2.974823	5.949646	5.949646
8	47	49.612753	-2.612753	-5.559048936	5.559048936
9	45	47.509613	-2.509613	-5.576917778	5.576917778
10	54	52.076282	1.923718	3.562440741	3.562440741
11	51	49.7368	1.2632	2.476862745	2.476862745
12	51	47.870002	3.129998	6.13725098	6.13725098
				平均	4.27232197

ここで, 表2の胸囲による推定値の相対誤差(15.60%), および, 表3の身長による推定値の相対誤差(5.07%)と比較すると, 相対誤差の平均(4.27%)が小さくなっており, 精度が向上していることが分かる.

3. 演習課題

表5に示した各都道府県の最高気温の月平均データについて, 例題の手順にしたがって分析しなさい. (最高気温の月平均を y として分析を行う. 「演習データ.xlsx」を利用すること).

表5 各都道府県の最高気温の月平均データ (一部抜粋)

都市名	緯度	月間降水量	最高気温の月平均
札幌	43.05	116.31	30.49
青森	40.82	131.38	31.81
秋田	39.72	172.31	33.15
...
宮崎	31.9	319.31	35.49
鹿児島	31.55	222.38	37.13
那覇	26.2	294.69	35.9

なお, レポートには, 表, 散布図(2つ), 単回帰式(2つ), 相関係数(2乗のままでよい, 2つ), 単回帰式による残差と相対誤差の表(2つ), 重回帰式, 重相関係数, 重回帰式による残差と相対誤差の表を載せること.

4. 実験課題

図 1 6 に示したように電気情報工学科棟には実験用にいくつかの無線 LAN アクセスポイントが設置されている。体育館側を 0m とし、中央の階段付近で 50m である。これらのアクセスポイントからの電波強度 (RSSI 値) を測定することで、現在、どこの場所にいるのかを特定することができる。

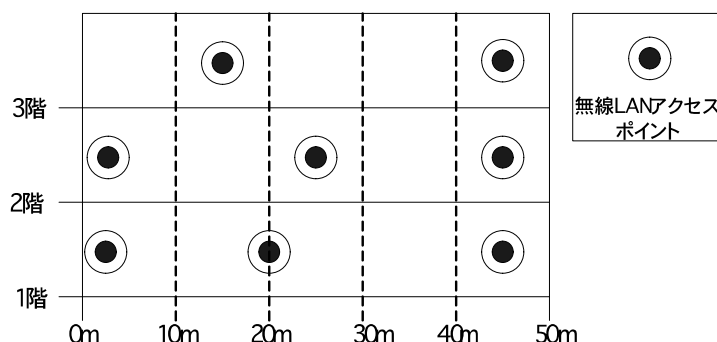


図 1 6 : 電気情報工学科棟の実験用無線 LAN アクセスポイント

4. 1. 実験方法

電波強度の測定にはノートパソコンを利用する。5 m 間隔ごとにアクセスポイントからの電波強度を測定する。1 階から 3 階までのどの階を測定するかは、班毎に指定する。

測定には、実験用に特別に作成したソフトウェアを利用する。各測定ポイントにパソコンを固定し、その場所の階数と何メートル地点かを指定し、ソフトウェア上の「測定ボタン」を押すことで、一度に 5 セットのデータを収集することができる。

表 6 : 収集される電波強度のデータ

位置	AP1	AP2	AP3	AP4	AP5	AP6	AP7	AP8
100	-46	-55	-54	-61	-80	-80	-80	-80
105	-53	-46	-54	-59	-80	-80	-80	-80
...								
225	-80	-69	-80	-59	-31	-57	-80	-80
230	-80	-62	-80	-58	-65	-47	-77	-70
...								
345	-80	-80	-68	-80	-78	-54	-55	-34
350	-80	-80	-72	-80	-80	-59	-63	-44

収集したデータは、表 6 のようなデータである。最初のカラムの 3 桁のデータは、測定地点の位置のデータであり、100 の位が階数を意味しており、のこり 2 桁で何メートル地点かを表している。以降に続くデータは、それぞれのアクセスポイントの電波強度の値である。実際に出力されるファイルは、カンマ区切りの CSV 形式のファイルで出力される。CSV 形式のファイルは Excel で開くことができるので、そのまま重回帰分析を行うことが可能である。

4. 2. 実験 1

指定された階において、0m, 5m, 10m, ..., 50m と 5 m 間隔で 5 セットのデータ (合計 55 個のデータ) を収集する。収集したデータを利用して、位置と電波強度における重回帰分析を行う (手順 5)。重回帰分析の結果から重回帰式および重相関係数を導出する (手順 6)、最後に、求めた重回帰式より推定値を計算し、相対誤差の絶対値の平均値を求める (手順 7)。

4. 3. 実験2

実験1と同様の手順で行う。ただし、重回帰分析は、実験1および実験2の両方のデータを利用して行う。

5. 考察課題

- 重回帰分析の原理（説明、解き方など）についてまとめ、本実験レポートの原理として記述しなさい。
- RSSI 値とはどのような値かを調べなさい。
- 重回帰分析におけるデータ数と精度の関係について考察しなさい。
- 相関係数だけでデータの関係性を判断することはできない理由を考察しなさい。
- 身の回りにある事柄やデータについて、データマイニングを応用することで有効な情報が得られる事例について考察しなさい。

○レポートの構成

①目的

②原理

重回帰分析の原理

③演習課題

以下のものを載せること。

散布図（2つ）、単回帰式（2つ）、相関係数（2乗のままでもいい、2つ）、

単回帰式による相対誤差の絶対値の平均（2つ）、

重回帰式、重相関係数、重回帰式による相対誤差の絶対値の平均。

④実験課題

実験方法

実験1の結果（重回帰式、重相関係数、相対誤差の絶対値の平均）

実験2の結果（重回帰式、重相関係数、相対誤差の絶対値の平均）

⑤考察

考察課題

⑥感想

⑦参考文献