重回帰分析への質的データの投入について

一般に、重回帰分析は量的な目的変数を量的な説明変数で予測する手法で、数量化Ｉ類は量的な目的変数を質的な説明変数で予測する手法であると言われている。しかし、重回帰分析の説明変数には量的データだけでなく質的データも使うことができ、質的データだけの場合、その結果は数量化Ｉ類の結果と一致する。以上のことを実際のデータを使って調べてみる。ここでの分析は、メニュー［分析－多変量解析他－予測手法－］の中の重回帰分析と数量化Ｉ類を使っている。

図 1 のデータは、大学成績を高校成績、勉強時間、出席状況、アルバイトの有無で予測しようとしたものである。ここに出席とアルバイトは質的なデータである。


データ編集 統計分析の意味と関連性.txt

	大学成績	高校成績	勉強時間	出席	アルバイト	出席1	出席2	出席3	出席4	アルバイト1	アルバイト2
1	72	4.6	3.7	1	1	1	0	0	0	1	0
2	88	3.4	5.7	2	1	0	1	0	0	1	0
3	67	4.0	3.1	2	1	0	1	0	0	1	0
4	76	3.2	2.4	2	2	0	1	0	0	0	1
5	95	5.4	5.3	4	2	0	0	0	1	0	1
6	64	3.8	2.1	1	2	1	0	0	0	0	1
7	68	4.0	2.2	1	1	1	0	0	0	1	0

2/6 (1,1)　分析:　備考:

図 1　データ

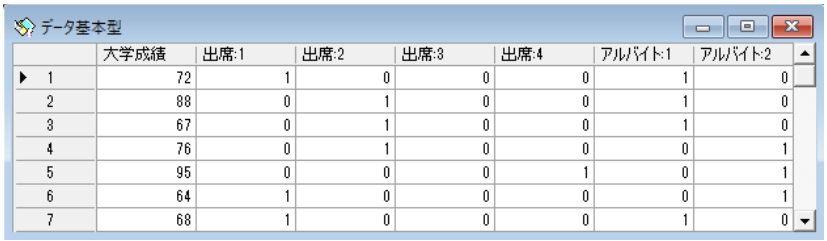
まず大学成績を目的変数に、高校成績、勉強時間、出席、アルバイトを説明変数にした重回帰分析の結果を図 2 に示す。但し、出席とアルバイトは質的データをそのまま量的データのように使っている。右下の青い部分は結果表示では別に表されるが、重要なので加えている。


偏回帰係数と検定

	偏回帰係数	標準化係数	t検定値	自由度	確率値	相関係数	偏相関係数
高校成績	1.6560	0.1933	2.0609	40	0.0459	0.398	0.310
勉強時間	3.4425	0.4989	4.5705	40	0.0000	0.693	0.586
出席	3.1429	0.3482	2.9736	40	0.0050	0.729	0.425
アルバイト	4.0901	0.2233	2.5932	40	0.0132	0.175	0.379
切片	42.9551	0.0000	10.1941	40	0.0000	R^2	0.716

図 2　順序尺度をそのまま利用した重回帰分析結果

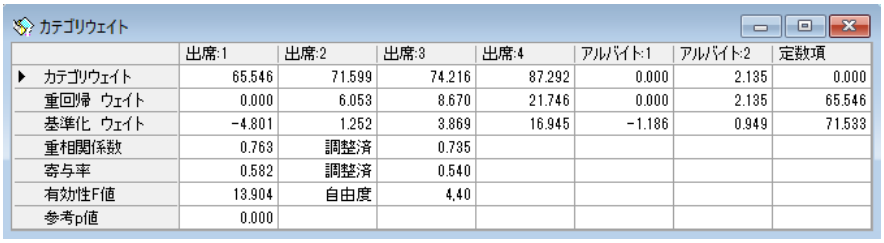
次に大学成績を目的変数に、出席とアルバイトを説明変数にして、数量化Ｉ類の方法で予測式を作ってみる。数量化Ｉ類ではデータを図 3 のような 0/1 データに変換して重回帰分析を実行する。その係数をカテゴリウェイトという。


データ基本型

	大学成績	出席:1	出席:2	出席:3	出席:4	アルバイト:1	アルバイト:2
1	72	1	0	0	0	1	0
2	88	0	1	0	0	1	0
3	67	0	1	0	0	1	0
4	76	0	1	0	0	0	1
5	95	0	0	0	1	0	1
6	64	1	0	0	0	0	1
7	68	1	0	0	0	1	0

図 3　数量化Ｉ類の計算のための形式

ここでは変数をアイテム、その中の分類をカテゴリという。分析結果を図 4 に示す。



カテゴリウェイト

	出席:1	出席:2	出席:3	出席:4	アルバイト:1	アルバイト:2	定数項
▶ カテゴリウェイト	65.546	71.599	74.216	87.292	0.000	2.135	0.000
重回帰 ウェイト	0.000	6.053	8.670	21.746	0.000	2.135	65.546
基準化 ウェイト	-4.801	1.252	3.869	16.945	-1.186	0.949	71.533
重相関係数	0.763	調整済	0.735				
寄与率	0.582	調整済	0.540				
有効性F値	13.904	自由度	4,40				
参考p値	0.000						

図 4　数量化 I 類分析結果

数量化 I 類では、カテゴリウェイトの値がカテゴリ間の制約条件（合計が 1 になる）から一意的に定まらず、ここでは値が 3 種類の形式で与えられている。最も結果の解釈が容易なカテゴリウェイトは基準化カテゴリウェイトである。この 0/1 形式のデータを用いて重回帰分析を行った結果と一致するのが、重回帰カテゴリウェイトである。但し、重回帰分析には、上に述べた制約条件から多重共線性という問題が生じ、すべての 0/1 データを使うことができない。そのため各アイテムの先頭カテゴリ（もちろん他のカテゴリでもよい）を除いて分析を実行する。分析結果を図 5 に示す。

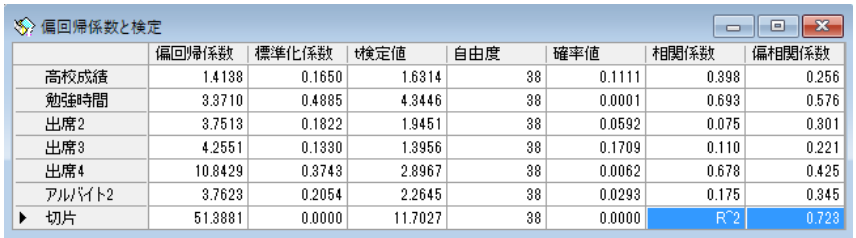
偏回帰係数と検定

	偏回帰係数	標準化係数	t検定値	自由度	確率値	相関係数	偏相関係数
出席2	6.0529	0.2940	2.7341	40	0.0093	0.075	0.397
出席3	8.6699	0.2710	2.5247	40	0.0156	0.110	0.371
出席4	21.7458	0.7507	7.0178	40	0.0000	0.678	0.743
アルバイト2	2.1348	0.1165	1.1015	40	0.2773	0.175	0.172
▶ 切片	65.5463	0.0000	38.4690	40	0.0000	R^2	0.582

図 5　質的データを用いた重回帰分析結果

この結果の偏回帰係数を数量化 I 類の重回帰カテゴリウェイトと比較すると同じものであることが分かる。

最後に、再び説明変数に量的データと質的データを混ぜることを考える。数量化 I 類の手法を用いるのであれば、0/1 データで各アイテムの先頭カテゴリを削除して重回帰分析を実行する。



偏回帰係数と検定

	偏回帰係数	標準化係数	t検定値	自由度	確率値	相関係数	偏相関係数
高校成績	1.4138	0.1650	1.6314	38	0.1111	0.398	0.256
勉強時間	3.3710	0.4885	4.3446	38	0.0001	0.693	0.576
出席2	3.7513	0.1822	1.9451	38	0.0592	0.075	0.301
出席3	4.2551	0.1330	1.3956	38	0.1709	0.110	0.221
出席4	10.8429	0.3743	2.8967	38	0.0062	0.678	0.425
アルバイト2	3.7623	0.2054	2.2645	38	0.0293	0.175	0.345
▶ 切片	51.3881	0.0000	11.7027	38	0.0000	R^2	0.723

図 6　0/1 データに変換した順序尺度を用いた重回帰分析結果

ここで、図 2 の順序尺度をそのまま使った重回帰分析の結果とこの数量化 I 類の方法を使った結果を比較してみよう。前者の寄与率が 0.716 であるのに対して、この方法の寄与率は 0.723 と向上している。しかし、データ利用の便利さを考えると、前者の方法を使うことはそれほど悪いことではない。ただ、質的データが順序尺度でなく名義尺度の場合は、0/1 データに変換する方法を使う必要があるだろう。また、アルバイトのように 2 つのカテゴリの場合は、どちらの方法でも結果は同じである。最後に、この考え方は判別分析やその他の分析の場合にも適用できることを付け加えておく。