茨城キリスト教大学紀要第43号　247
自然科学　p.247～262

教育統計学の講義研究（２報）

土　井　　努*

はじめに

本稿は筆者が受持っている教育統計学Ⅱにほぼ該当する講義内容と，その際の注意点などをまとめたものである。講義は1997年から継続しているものの，内容は当初から大幅に変わり，現在はここに報告するような形となった。本報告の主題は相関および回帰であるが，これらは統計学の中でも最も実用性の高い分野であるため，よくありがちな誤解，例えば相関係数は直線の傾きを表すなどの誤解を防ぐ事にも重点を置いている。

対象が文科系の学生であるため，授業では数式についての説明はほとんど省く一方，具体例によって手順を示し，次に実データを用いて演習を行っている。データの分析は手軽さの点から，Excelを用いているが，提出すべきレポートにおいては不十分でも，学生自身の直感と解釈によって書くよう強調している。これは便利な統計ソフトが普及した現在，結果の解釈が追いつかない傾向となりがちな事に対応するためある。

最終節「理論的な検討」は，授業では部分的に取上げている節であるが，より深く学習する時のため，例題を用い理論的背景を全般にわたって述べた。

１．相関分析

１報では，対応のあるデータ（データの組）がクロス集計表の形式であるときの関連度について，次の３種類に分け，相互に比較可能な相関を考えた。

・データの組が名義尺度（物の名前で順序が無い質的変量）のときCramerの属性相関
・データの組が順序尺度（数値でないが順序のある変量）のときspearmanの順序相関
・データの組が間隔尺度（普通の数値変数）のときピアソンの積率相関

ここでの目的は，データの組が間隔尺度である一方，クロス集計表とは限らない一般の場合について，ピアソンの積率相関（普通，相関係数と呼ばれている）を用いて，関連性を調べる事である。関連性（または相関）と因果関係とは，互いに別の概念であり，社会データにおいては関連性はあるが，どちらが原因または結果と決められないものも多く存在する。統計的な方法は多数データの傾向を探る手法なので，因果関係についてはより強い条件が必要な事を念頭に以下の分析を進めたい。

*茨城キリスト教大学　文学部　非常勤講師

1.1　相関係数の定義と意味

講義では相関係数の定義式に入る前に，まず実データを適切な散布図に表す練習から入り，これによってデータ間にはどのような関連性が存在するかなど，社会的な背景を含めた検討から始めている。この段階においてExcelによる作図力が意外と問われる事が多いので，パソコンの話になるが，問題点をまとめてみると，

・散布図（縦軸，横軸ともに数値）の作成自体を会得していない事が多い。
・元データに対する，グラフの縦横軸との対応付けが意識されていない場合があるで，作成した図の軸の意味が曖昧に成り易い。（グラフ軸の名前は自動的に入らない。）
・目盛を適切に設定しないために点が一箇所にまとまることが良くある。分析データを視覚的に良く観察する事は，統計手法に入る前段階として非常に大切と思われる。

さて，散布図が自分の期待する程度に描け，組データの関連性について，各自の解釈が出た段階で，相関係数の定義式へと，またこれを計算する為のExcel命令へと進む。

相関係数 r ＝ x，y の共分散／x の標準偏差／ y の標準偏差
＝Σ｛(X−AVx)(Y−AVy)｝／(N−１)／SD x／SDy

ここで

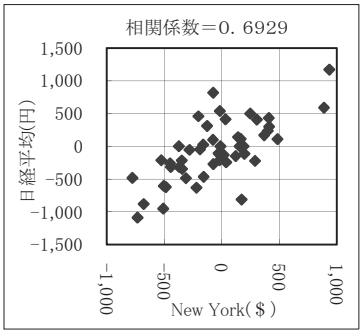
x，y の共分散　：Σ｛(X−AVx)(Y−AVy)｝／(N−１)
x の標準偏差^2　：$(SD\ x)^2$＝Σ｛$(X-AVx)^2$｝／(N−１)
y の標準偏差^2　：$(SDy)^2$＝Σ｛$(Y-AVy)^2$｝／(N−１)
X　：組データのうち散布図の x 軸とするデータ
Y　：組データのうち散布図の y 軸とするデータ
AVx：データXの平均
AVy：データYの平均

また，組データの相関係数を計算するExcel命令は，次の統計関数による。

CORREL（xの範囲，yの範囲）

図表1.1は１例として組データと，その散布図を示したものであるが，参考用に相関係数の値を統計関数で計算の後，図に記入してある。

日	X:NewYork($)	Y:日経平均(円)
11／10	−73.27	−272.13
11／7	248.02	498.43
11／6	−443.48	−316.14
11／5	−486.01	−622.1
11／4	305.45	406.64
11／3	−17.1	537.62
10／31	156.24	NA
10／30	189.73	NA
10／29	−74.16	817.86
10／28	889.35	589.98
10／27	−203.18	459.02
・	・	・
9／4	−344.65	−345.43



図表1.1　New York，東京株式市場の対前日値幅（2008.Sep.−Nov.）

1.2　相関係数の性質と注意点

授業では以下のように，相関係数についての一般的な性質を述べた後，多くの学習者が誤解しやすい点（＊項目）について，特に注意している。

・相関係数は，組データ（x，y）の直線的な関係の度合いを，数値で表現したものである。
・相関係数は－１から＋１の間の小数値を取る。
・正の相関係数とは，全体的な傾向として，ｘとｙは**正**比例的である事を言う。
・負の相関係数とは，全体的な傾向として，ｘとｙは**逆**比例的である事を言う。
・相関係数の絶対値が１に近い時ｘとｙはほとんど直線的な関係にある事を示す。
・相関係数の絶対値が０に近い時xとyは直線的な関係をほとんど持たない事を示す。
・相関係数の値は，原点のとり方によって，変化する事は無い。
＊相関係数の値は，直線の傾きとは無関係である。なぜならば直線の傾きは単位によって変化するからである。
＊相関係数は，組データ（x，y）間の因果関係を必ずしも表さない。相関関係と因果関係とは独立した別の概念である。

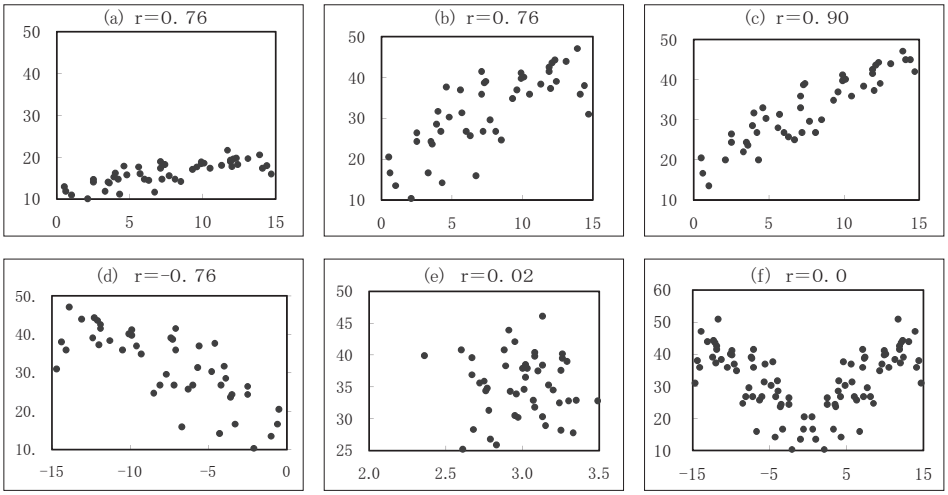
図1.2に散布図を用いて，相関係数の性質と注意点とを明らかにすると，

⑷に対して⑹の傾きは急になっているが，相関係数は変わらない。その理由は，縦軸の単位を⑹は一般的な１次式による単位変換の一例，$y_b = 3.5\, y_a - 25$によっただけである事による。このように点の傾きは単位変換に依存するが，相関係数は不変である。

⑹は相関係数0.9の例であり，⑹に比べて直線的である。⑷～⑹より相関係数は，直線回りの点の集中度合いを表す事がわかる。

⑺は負の相関の例，⑻は相関≒０となる現実のデータの例である。

⑼のように点が左右対称形のとき，直線近似は出来ないので，直線への集中度合いを表す相関係数は０と計算される。このように散布図を描く事は非常に重要である。



図表1.2　相関係数の例（性質と注意点）

1.3　相関係数が1.0付近の時

社会データを分析していると，一対のデータ（x，y）の散布図がほとんど直線に載ることがよくある。授業ではこのような場合の例を用い，原因と対策を学んでいる。

原因：組データ（x，y）が社会制度，経済原則などのため比例関係にあるとする。人口規模，経済規模によりデータが変化すると，点（x，y）が直線上を比例的に動く場合である。

対策：率x／yを一方の軸として，人口規模などの影響がでないようにする。また年次変化などが絡むものは，時系列的なx，yの増減率または増減幅をとる。

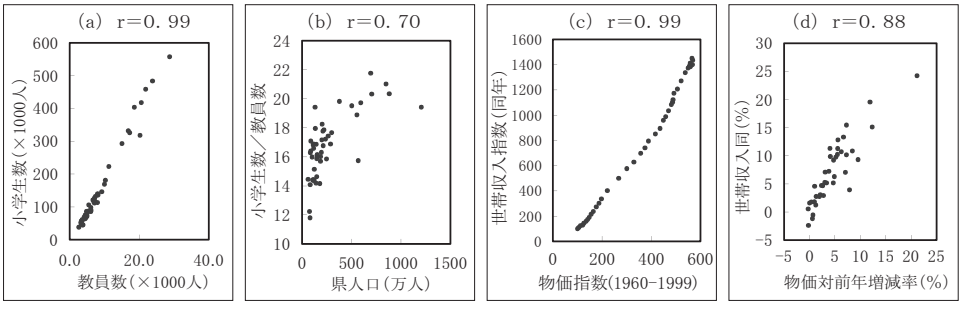
図表1.3に，組データが直線的な関係を示す例を挙げる。

⑷小学生数と教員数に関する，県別の組データを示したものであるが，両者の人数に関する関係は，学校制度上ほぼ決っているので，小学生数yと教員数xとは比例の関係，すなわち県の人口に依存して一直線上に乗る関係となる例である。すなわち県別の人口規模が隠れた第３の要因となっている例である。

⑹そこで組データの比を取ってみる。すなわち小学生数／教員数を縦軸に，そして横軸には第３の要因である県別人口をとってみると，制度上決っていた比例関係でも，県人口によるばらつきがある事が示された。

⑽1960年から40年間に渡って，物価指数と世帯収入を組データとしたときの変化を考える。この散布図は，先進国においては物価と世帯収入とは経済構造上，比例している事を，両者の直線関係が，改めて示しているものの，この散布図自体は，ばらつきを扱う統計データの対象とはなりにくい。すなわち経済的な比例関係が，年次変化によって直線的な図となった事が原因である。

⑾前の例では年次変化が，隠れた第３の要因となっていた。物価指数と世帯収入，各々につい前年度に対する増減率を考える。すると年度によって，両者の増減率は統計的にばらつく事がわかり，その理由など次の分析に進む事ができよう。



図表1.3　組データの相関係数が1.0に近い時

このように対象とするデータが特に人数，金額，個数などの素数のとき，組データの関係はほとんど直線的である場合が少なくない。その原因となっている社会制度や人口の影響，あるいは経済原則や年次変化の影響などを取り除くため，比率をとる事が有効な場合が多い。なお図1.1は率に代わって，差をとる事により適当な散布図を得ている。

1.4　見かけの相関について

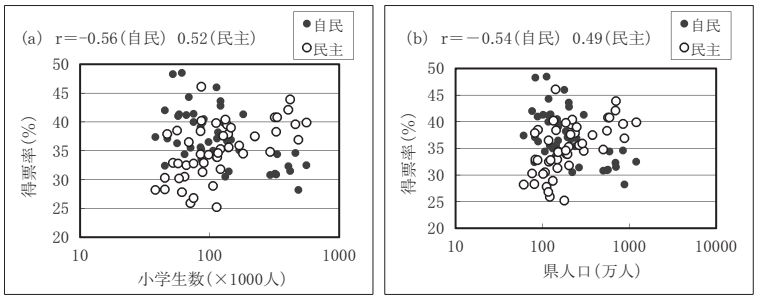
古典的なヨーロッパの笑い話「出生率が増加した年は，コウノトリの数も増えていた。それはコウノトリが赤ちゃんを運んでくるらしい。」は意味の無い見かけの関連を言ったものであるが，このように相関係数が十分に大きい場合でも，社会的に相関の理由が挙げられない場合があり，これを見かけの相関または擬相関と呼んでいる。相関に関する学習過程ではよく遭遇する現象なので，具体例を用いて学んでいる。見かけの相関があるときは別の要因が背後に隠れている事に注意し，この要因を探し出す必要がある。

図表1.4は2003年衆議院選挙における，２大政党の得票率を県単位に示している。

⑷この図では県別の小学生数をｘ軸に，得票率をｙ軸にとった。図をそのまま解釈すると，選挙権のない小学生数が多い県ほど民主党の得票率が上がり，自民党の得票率が下がるといった，説明のつかない図となっている。もちろん，県別の小学生数は県人口に最も左右されるので，県人口が見かけの相関を作り出した要因と考えられる。

⑸見かけの相関を作っていた要因である県人口を横軸にとった散布図を考える。これより2003年時点では，人口の多い県が民主党の支持率が高く，逆に人口の少ない県が自民党の支持率が高い傾向を持つ事が読み取れる。これは一般に言われていた概念である都市部は民主，農村部は自民との見方を反映したデータと解釈できるので，⑷のような見かけの相関が取り除かれ，社会的に説明のつく統計データとなった。

このように統計分析から明らかになるのは，データが持つ数字としての傾向だけなので，この傾向の理由は，因果関係など，一般的，社会的なメカニズムから説明される必要がある。もし，この説明ができない時は，見かけの相関関係を疑い，その要因を探し出すため，次節に示すように，相関係数行列による検討が必要である。



図表1.4　見かけの相関⑷と，本来の相関⑸

1.5　相関行列による検討

相関の演習用に，図表1.5に示すような県勢データを用いているが，ここで(a)は県別の各種データ，(b)はそれらの相関行列である。ここでは相関係数が1.0に近い場合や，見かけの相関などについて県勢データの中で，改めて検討してみよう。

相関行列の作り方は，Excel上で，ツール→分析ツール→相関　と選ぶ。

・1.3節で取上げた相関係数が1.0付近になる組データは，人口，小学生数，教員数の相互である事がわかるが（相関係数0.99），いずれも社会的にも関連性が強いことは明らかであろう。すなわち県別に見ると，世代構成のばらつき以上に，人口のばらつき（その幅は県によって約20倍）の方が大きいと言えるからである。演習では多くの組データの中から，このようなものを見つけ，その理由を明確にすることを目指している。

・1.4節で取上げた，見かけの相関は組データの中において数多くある事がわかる。前述の組合せ以外にも，平均月給，小学生数，自民，民主党得票率など相互の関連性は社会的な背景からは説明するのが難しいことが多いと言えよう。社会データは一般に，ある側面からは真であるが，別の側面からは偽であることが多いので，見かけの相関を見つけるのは場合により容易ではないが，社会背景を考える演習教材としてこのデータを用いている。

図表1.5　県勢データ

(a)県別のデータ

県名	人口（万人）	人口密度（人/㎢）	一次産業（%）	二次産業（%）	世帯人数（人）	65歳以上（%）	平均月給（万円）	小学生 ×1000人	小学教員 人	自民得票 %	民主得票 %
北海道	568	72	8.1	22.9	2.6	14.8	306	318	20208	31	40.8
青　森	148	154	14.7	25.6	3.07	16	285	90	6151	40.5	32.9
岩　手	142	93	14.4	30.2	3.13	18	299	86	6061	33.5	46.1
宮　城	237	325	6.3	26.9	3	14.5	333	139	8074	35.4	37.9
秋　田	119	102	12.1	30.9	3.24	19.6	314	66	4572	41.2	32.5
山　形	124	133	12.3	33	3.49	19.8	310	75	4950	41.4	32.8
福　島	213	154	10.5	34.2	3.26	17.4	312	136	8112	36.7	39.5
茨　城	299	490	8.5	33.8	3.2	14.2	337	181	10251	41.3	34.5
	・	・	・	・	・	・	・	・	・	・	・
沖　縄	132	581	6.7	18	3.15	11.7	277	106	5458	35	28.9

(b)相関行列

	人口	人口密度	一次産業	二次産業	世帯人数	65歳以上	平均月給	小学生	小学教員	自民得票	民主得票
人　口	1.0000										
人口密度	0.8759	1.0000									
一次産業	−0.6500	−0.5920	1.0000								
二次産業	−0.1126	−0.1534	−0.2511	1.0000							
世帯人数	−0.5733	−0.5212	0.2490	0.5732	1.0000						
65歳以上	−0.7137	−0.5786	0.7607	−0.0085	0.2549	1.0000					
平均月給	0.7366	0.7517	−0.7354	0.2516	−0.3788	−0.5082	1.0000				
小学生	0.9920	0.8358	−0.6550	−0.0912	−0.5455	−0.7512	0.6939	1.0000			
小学教員	0.9868	0.8064	−0.6138	−0.1290	−0.5792	−0.7100	0.6733	0.9915	1.0000		
自民得票	−0.5393	−0.3994	0.3007	0.2840	0.4761	0.4405	−0.3773	−0.5598	−0.5691	1.0000	
民主得票	0.4976	0.2894	−0.4422	0.3111	−0.0212	−0.5218	0.4820	0.5178	0.5186	−0.5291	1.0000

出典：データで見る県勢　（矢野記念会　2002，他）

２．回帰分析

回帰分析は統計手法のうちで最も広範囲に使われ，実用性が高い手法と見る事ができる。回帰分析で扱うデータは，数値でなければならないが，数値から質的なデータへ拡張された場合，回帰分析は数量化理論１類，２類，その他へと拡張される。この意味において，回帰分析の基礎的な概念を学ぶ事は非常に重要であり，教育統計Ⅱの後半に取り入れている。授業では将来，学生がさまざまな分析ソフトを利用した時に得られるであろう結果に対して，基本的な解釈と説明ができるようになる事を目指している。

授業で用いるデータは簡単な２要因の例で，計算はExcelの分析ツールを利用して教材としての手軽さと一般性とを保つよう留意している。

2.1　回帰分析への導入

回帰分析への導入部では中学，高校における１次関数の知識から延長して考えられるように，単回帰から，そして重回帰を説明している。

・単回帰

y＝a＋bxで表される１次関数によってyを推定する事を考える。

y：目的変数，すなわち分析対象とみなすデータ

x：説明変数，すなわちyに影響を及ぼす要因データ

係数a，bを計算するため，多数のデータの組（x，y）を統計ソフトに入力すると，y方向の誤差が最も少なくなる（x方向ではない）ように，a，bが決められる。

・重回帰

目的変数の値をyとするのは単回帰と同様であるが，複数の要因が影響を及ぼすと考えられるとき，これらの要因をx_1，x_2・・・とモデル化する。

そして　$y=b_0+b_1x_1+b_2x_2+\cdots$

で表される１次式によって，yを推定する事を考える。係数b_0，b_1，b_2・・を求める方法は多数のデータの組（x_1，x_2，y）を統計ソフトに入力する。

2.2　回帰分析の手順

回帰分析の手順を例題を用いて説明する。例題は，ある30人の年齢，血圧，肺活量を測定した結果が図表2.1(a)，データ間の相関行列が同図(b)のように示された時，血圧または肺活量から年齢を，単回帰または重回帰によって推定することを考える。

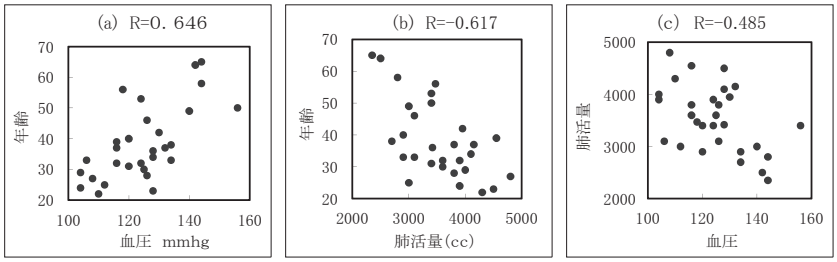
図表2.1　回帰分析用の医療データ

(a) 元データ

	X1	X2	Y
データno	血　圧	肺活量	年　齢
1	110	4300	22
2	128	4500	23
3	104	3900	24
・	・	・	・
30	144	2350	65

(b) 相関行列

	血圧	肺活量	年齢
血　圧	1.000		
肺活量	−0.485	1.000	
年　齢	0.646	−0.617	1.000



図表2.2　医療データ例の散布図

①分析目的に従い，データを目的変数yと説明変数xとに分け，変数間の相関係数および散布図を描く。このとき飛び離れた点や，1箇所に固まった点が多いときは注意が必要であるので，その原因を抑えておく。

・例題は，目的変数yは年齢，説明変数xは血圧または肺活量とモデル化される。まず相関行列を視覚化するため，変数間の散布図を図表2.2のように作る。

②目的変数と相関の高い説明変数xを選ぶ。

・血圧は年齢と0.646の相関があり，また肺活量は年齢と−0.617の相関がある。この例では他の説明変数がないので，同符号を持つ相関係数の相対比較はできない。また上述の説明変数は年齢に対し互いに逆の傾向を持つので，共に適当であると判断する。

③重回帰における説明変数は，②の条件に加え，互いに相関の低いものを選ぶ。もし，互いに相関が高いならば，説明変数は片方で十分，あるいは寄与度が低いとみなす。

・例題では，血圧，肺活量間の相関係数は，−0.481であり，この値は他の相関係数より絶対値が小さいだけでなく，負なので互いに逆の傾向を持っている事がわかる。よってこれらは重回帰分析の説明変数として組み込む必要があると判断できる。

④元データを次式によって平均0，標準偏差1となるように標準化する（図表2.3）。これはデータの単位によらず，回帰係数が比較できるようにする為である。

標準化したデータ＝(x−AV)／SD

xは元のデータ，AVはその平均，SDは標準偏差

図表2.3　元データを標準化

	← 元データ		→	← 標準化データ		→
	X1 血圧	X2 肺活量	Y 年齢	X1 血圧	X2 肺活量	Y 年齢
1	110	4300	22	−1.15	1.22	−1.33
2	128	4500	23	0.26	1.54	−1.25
3	104	3900	24	−1.62	0.58	−1.16
・	・	・	・	・	・	・
30	144	2350	65	1.51	−1.91	2.24
AV	124.7	3539.7	38	0	0	0
SD	12.772	623.1	12.047	1	1	1

⑤回帰分析の実行は，以下の命令を行うことによる。

Excelにおいて，ツール→分析ツール→回帰分析

・例題では，標準化したデータに対して，単回帰および重回帰を行った結果を次にまとめた。なお，データが標準化されているので，定数項はどの場合も０となる。

血圧を用いた単回帰：　年齢y＝０＋0.6462×血圧
肺活量を用いた単回帰：　年齢y＝０－0.6172×肺活量
血圧と肺活量を用いた重回帰：　年齢y＝０＋0.4535×血圧－0.3972×肺活量

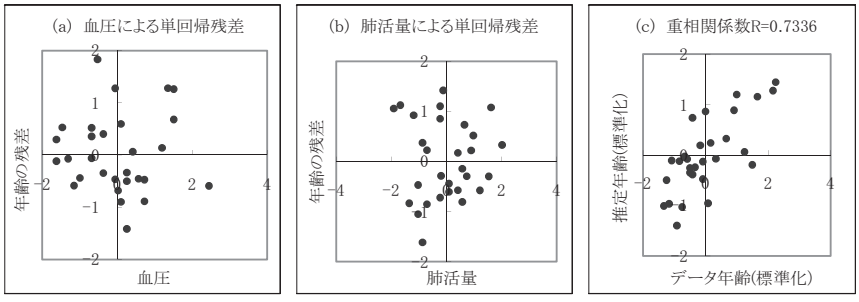
⑥回帰に用いたデータが標準化されているので，回帰係数の比較を行う事が出来る。大きい回帰係数に対応する説明変数の方が，要因としての寄与度が大きいと言える。

・回帰係数の絶対値の比較により，説明変数として，血圧の方が肺活量より寄与度が多少大きい事がわかる。なお単回帰における回帰係数は相関係数に等しい事がわかる。

⑦回帰推定の結果を調べるため，残差（元データyと推定値との差）の散布図を描き，次の条件が満たされている事を確認する。

残差は説明変数の大きさに依存しない。
経験的には大多数の残差が±1.5または±2.0の中に納まる。
重回帰では目的変数について，推定値と元データ間の相関係数が大きいこと。

・例題では，血圧を用いた単回帰の残差を図表2.4(a)に，肺活量を用いた単回帰の残差を同図(b)に示た。重回帰については残差の代り，年齢の推定値と元データyとを同図(c)に示した。これによって，両者の相関係数を視覚化した。図の(a)～(c)どれについても，上述の条件が満足されていることが確認できる。



図表2.4　回帰推定における残差など

⑧回帰推定が有効であるのは，元データの最小値から最大値までの範囲内とするのが原則となる。この理由はデータの範囲外について，x，yは直線的な傾向を持つか否か，推定できないからである。反面，統計以外の知識から，x，yの直線的な傾向が保障されるならば，直線を延長する事は可能なように思えるが，回帰係数が従来と同一であるとは保障されていないので注意が必要である。

2.3　理論的な検討

これまでは回帰分析の結果を残差の散布図などによって，視覚的に検討してきたが，仕上げとしてExcel出力の解釈について学ぶ。なお重要な項目（＊印）を優先的に解説し，数多い理論的な項目全体にメリハリを付けた。以下の検討はこれまでと同じ例題についての次の３ケースである。以下では　n：データ数，p：説明変数の個数。

a.　血圧を用いた単回帰：　年齢y＝０＋0.6462×血圧

b.　肺活量を用いた単回帰：　年齢y＝０－0.6172×肺活量

c.　血圧と肺活量を用いた重回帰：　年齢y＝０＋0.4535×血圧－0.3972×肺活量

2.3.1　回帰統計の表

Excelの出力中，回帰統計の表を図表2.5に示すが，結果の数値は，データが標準化されていない場合でも同じとなる。

図表2.5　回帰統計の表

(a)	
重相関R	0.6462
重決定R2	0.4176
補正R2	0.3968
標準誤差	0.7767
観測数	30

(b)	
重相関R	0.6172
重決定R2	0.3810
補正R2	0.3589
標準誤差	0.8007
観測数	30

(c)	
重相関R	0.7336
重決定R2	0.5382
補正R2	0.5040
標準誤差	0.7043
観測数	30

＊重相関　R：目的変数yについてデータ値と回帰推定値（回帰式にxの値を代入して計算）との相関係数である。この値が大きいほど，回帰推定が良い事を示す。

a：0.6462（単回帰の時は相関係数の絶対値と等しくなる）

b：0.6172（　　〃　　）

c：0.7336（重回帰によって，重相関は単回帰の時より向上している）

＊重決定R^2：重相関の自乗として計算される。その意味は，目的変数yの全変動(2.3.2参照）のうち，回帰による変動割合を示すので，寄与率とも呼ばれる。上述の重相関Rは回帰の有効性について，相関係数の面から評価したのに対し，ここでは変動の面から回帰が寄与する割合を評価している。

a：$0.6462^2=0.4176$

b：$0.6172^2\fallingdotseq 0.3809$

c：$0.7336^2=0.5382$

・補　正　R^2：重決定係数R^2の値は説明変数を多くすれば向上するが，説明変数は実用上，必要最小個数である事が望ましい。そこで変数が多くなると重決定係数R^2が小さくなる傾向に補正したものであり，次式で計算される。

$$1-(n-1)/(n-1-p)*(1-R2)$$

a：1－(30－1)／(30－1－1)＊(1－0.4176)＝0.3968

b：1－(30－1)／(30－1－1)＊(1－0.3809)＝0.3589

c：1－(30－1)／(30－1－2)＊(1－0.5382)＝0.5040

・標 準 誤 差：目的変数 yについて，データ値と回帰推定値との差を残差というが，標準誤差とは残差の標準偏差，(あるいは残差分散の√) のことである。別の見方をすれば推定誤差のばらつきとも言えるので，この値は小さいほど望ましい。なお下の確認計算にあたっては　分散分析表中の残差分散を用い，一般的性質である，標準偏差＝√分散　によった。

a：√(0.6032)＝0.7767

b：√(0.6411)＝0.8007

c：√(0.4960)＝0.7043

2.3.2　分散分析の表

Excel出力のうち，分散分析の表を図表2.6に示すが，ここでの目的は回帰が全体的に有効かを検討するものであり，その考えは回帰変動が誤差に埋もれないかを調べるものである。直感的に無理な回帰をしない限り，有効性は肯定されることが多い。

図表2.6　分散分析の表

(a)	自由度	変　動	分　散	観測された分散比	有意F
回帰	1	12.109	12.109	20.073	0.0001
残差	28	16.891	0.603		
合計	29	29			

(b)	自由度	変　動	分　散	観測された分散比	有意F
回帰	1	11.048	11.048	17.232	0.0003
残差	28	17.952	0.641		
合計	29	29			

(c)	自由度	変　動	分　散	観測された分散比	有意F
回帰	2	15.608	7.804	15.733	0.0000
残差	27	13.392	0.496		
合計	29	29			

・自由度：計算対象となるデータ数－関係式の数，すなわち独立に変化することのできるデータの個数のことである。

合計の自由度＝回帰用データ数－１＝30－１

回帰の自由度＝説明変数の個数

残差の自由度＝合計の自由度－回帰の自由度

・変　動：一般に，変動＝自由度×分散　の性質があり，合計の自由度は30－１だから，

合計(y)変動＝(30－１)×１　なぜなら標準化データではｙの分散＝１

また，残差変動＝合計変動－回帰変動

寄与率＝回帰変動／全体の変動

a：12.109(回帰変動)　16.891(誤差変動)　12.109／29＝0.4176(寄与率)

b：11.048　17.952　11.048／29＝0.3810

c：15.608　13.392　15.608／29＝0.5382

寄与率は変動の面から見て，回帰が寄与する割合を表すので，例においては単回帰よりも重回帰の方が良いことを示している。しかし寄与率は全体の半分程度なので，想定外変動が残りの半分を占めている事に注意する必要がある。

・分　散：分散＝変動／自由度　なる関係を回帰と残差各々について適用すれば，

a：12.109／１＝12.109(回帰)　16.891／28＝0.603(残差)

b：11.048／１＝11.048　17.952／28＝0.641

c：15.608／２＝7.8034　13.392／27＝0.496

・分散比：回帰分散／残差分散　によって定義される値で，回帰分散が残差分散の何倍あるかの値。これより，回帰は残差に埋もれないかの検定に用いる値となる。

a：12.109／0.603＝20.07

b：11.048／0.641＝17.23

c：７.804／0.496＝15.73

＊有意Ｆ：自由度（p，n－１－p）のF分布において確率変数が，上述の分散比以上となるような確率，すなわち回帰が残差に埋もれるような確率を表す。分散比が大きいほど，残差に埋もれる確率は下がることになる。確率の目安は一般に５％または１％を限界としているが，実用的にはこれより十分に小さい事が望ましい。この条件が満たされた時，回帰全体は有効であると推定するので，上述の確率は危険率と解釈できる。例題においてはどの危険率も0.1%以下で十分に小さく，回帰は残差に埋もれていないと見ることが出来る。

a：分散比＝20.07以上の確率＝0.0001

b：　〃　＝17.23以上の確率＝0.0003

c：　〃　＝15.73以上の確率＝0.0000

2.3.3　回帰係数の表

回帰係数に関するExcelからの出力を，図表2.7にまとめる。回帰係数は，当該データからの計算値なので，何度もデータを取り直せば，回帰係数自体ばらつきを持つと考えられ，これを予想するために回帰係数の表を用いる。

図表2．７　回帰係数の表

(a)	回帰係数	標準誤差	t	P－値	下限95%	上限95%
切　片	0.0000	0.1418	0.000	1.0000	－0.2905	0.2905
血　圧	0.6462	0.1442	4.480	0.0001	0.3507	0.9416

(b)	回帰係数	標準誤差	t	P－値	下限95%	上限95%
切　片	0.0000	0.1462	0.000	1.0000	－0.2995	0.2995
肺活量	－0.6172	0.1487	－4.151	0.0003	－0.9218	－0.3126

(c)	回帰係数	標準誤差	t	P－値	下限95%	上限95%
切　片	0.0000	0.1286	0.000	1.0000	－0.2638	0.2638
血　圧	0.4535	0.1496	3.032	0.0053	0.1466	0.7604
肺活量	－0.3972	0.1496	－2.656	0.0131	－0.7041	－0.0903

＊回帰係数：回帰分析で最も知りたいのがこの回帰係数である。なお例ではデータを標準化してあるので，切片は０となる。一般に単回帰の係数は相関係数と等しい。

a：y＝０＋0.6462×血圧

b：y＝０－0.6172×肺活量

c：y＝０＋0.4535×血圧－0.3972×肺活量

これより血圧は年齢と正比例する要因であり，係数が大きいので説明力が大きい事，肺活量は年齢と逆比例する要因であり，説明力は血圧より少し小さい事がわかる。このような回帰係数の比較は，標準化されたデータの時だけ可能であり，生データにおいては測定単位の影響が出るので，比較できない。

・標準誤差：前述のように回帰係数はデータによって，ばらつくと考えられる。そこで回帰係数の標準偏差（SD）を標準誤差と呼んで，その推定値を表に示してある。例において血圧，肺活量に対する回帰係数をa_1 a_2とすれば，

a：a_1のSD＝0.1442

b：a_2のSD＝0.1487

c：a_1のSD＝0.1496　　a_2のSD＝0.1496

・t 分 布：データから計算された回帰係数a_0 a_1 a_2は推定値であり，これらは，それぞれの真値（未知であるが）を平均として，標準誤差をばらつきとする確率変数と見なすことができる。そこで，次のように標準化した変数tを考える。

t＝（回帰係数の推定値－真値）／標準誤差

tは自由度n－１－pのt分布に従うことが分っている。この性質を利用して説明変数の有効性（回帰係数が０でない事に置き換える）を次のように求める。

・仮説検定：説明変数の有効性を言うためには，次のような論理展開が必要となる。

①仮説「回帰係数の真値は０である」を立てる。一方，対立仮説「回帰係数の真値は０でない」も同時に立て，両者で想定する全ての場合を尽すようにする。

②仮説によれば，回帰係数の真値＝０であるから，前項よりt値が計算できる。

t＝回帰推定値／標準誤差

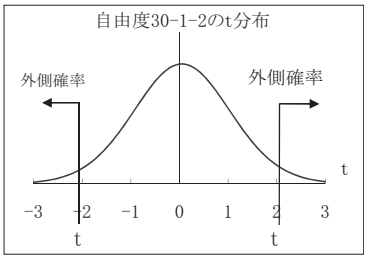
a：血　圧：　0.6462／0.1442＝　4.480

b：肺活量：－0.6172／0.1487＝－4.151

c：血　圧：　0.4535／0.1496＝　3.032

肺活量：－0.3972／0.1496＝－2.656

図表2.8　t分布の例



③t分布においてt値の大小により，次のように判断する。（図表2.8）

t値があまり大きくない→t分布の内側→「仮説通り回帰係数は０らしい。」

t値が十分大きい→t分布の外側→仮説棄却→「回帰係数は０でない。」

＊P値：「回帰係数が 0 でない」と判断した時でも，そうでない確率は，t 分布の外側確率だけあるので，この確率を危険率P値としている。危険率は十分に小さい必要があるので，通常 5 %以下を用いる。例題における危険率は，

P(t値，自由度)＝t分布における±t値より外側の確率

a：P(　4.48，28)＝0.0001　　血　圧は危険率0.01%を許容し有意と判断
b：P(−4.15，28)＝0.0003　　肺活量は危険率0.03%を許容し有意と判断
c：P(　3.03，27)＝0.0053　　血　圧は危険率0.53%を許容し有意と判断
　P(−2.66，27)＝0.0131　　肺活量は危険率1.3%を許容し有意と判断

＊注意：例えば「ケースa.血圧の回帰係数0.6462の値は，危険率0.01%で有意」であると誤解しないよう注意が必要である。検定の論理からは「回帰係数の値は 0 ではない」としか判断できず，その値は回帰係数の信頼区間によって求める。

＊注意：P値を有意確率と呼ぶ事もあるが，「説明変数が有意である確率」との誤解を招きやすい。そのように呼ぶと，有意確率なら大きい方が望ましいなどと誤解が広がってしまう。

＊上下限95%：回帰係数は，ばらつくことは述べたが，多数回の分析を行った時を想定して，回帰係数の取りうる範囲を95%の信頼度で示している。計算方法は，

表の回帰係数±標準誤差×内側95%のt値

a：血　圧　　0.6462±0.1442×2.0484＝下限　0.3587　　上限　0.9416
b：肺活量　−0.6172±0.1487×2.0484＝下限−0.9218　　上限−0.3126
c：血　圧　　0.4535±0.1496×2.0518＝下限　0.3069　　上限　0.7604
　肺活量　−0.3972±0.1496×2.0518＝下限−0.7041　　上限−0.0903

下線の欄はExcel（2000）の計算が間違っているので要注意。
ケースcにおいて，血圧の回帰係数がケースaより正確に，（上下限の幅，すなわち信頼区間が狭く）なった事により，重回帰を行った意味が表れている。
この表が示すようにP値では説明変数の有効性（回帰係数が 0 でない事）しか分らなかったが，95%信頼区間によれば回帰係数の範囲を推定できる。

この例が示すように，限られた 1 回のデータ（例では30個）から求めた回帰係数の値は，直感的に予想されるより，かなり広い推定幅を持つことがわかり，あまり正確とは言えない。より正確に推定を行うには，データ数を増す事が，数の傾向を扱う統計的な方法となる。しかし本質的には，個々のデータの信頼性を増すのが望ましいことは言うまでもない。

おわりに

この報告書と講義の進行スタイルは，一般の教科書とは逆の進め方をとってきた。すなわち，まず具体例を示し，これを直感的，視覚的に捉えた後，統計理論の説明を行う形である。特に力を注いだ点は，“相関係数が1.0付近の時”および“見かけの相関”そして“回帰における危険率”であり，理解の手助けになってきた事を願っている。

１，２報で述べた内容は，ほぼ最近の講義内容と同じであるが，今後の授業を進める際の別案として，さらに狭めた内容を余裕を持って構成した方が，学習する側にとって，達成感が得られるのではないかと，レポートに書かれた感想から最近思う。講義で取り扱わない範囲についても，達成感を元に必要に応じて独学できるようにも感じる。

計算ソフトはExcelを利用してきたが，コンピュータ上では，とかく操作に神経が流れがちな事を加味すると，誤解しやすい点を注意表示するなどの独自のソフト開発の必要性を感じる事もある。しかし本来は，印刷物の上でじっくり考える習慣をつけることが先決のように思うので，授業における個々の対話をやはり重視して，小規模大学にふさわしい授業形態としていきたい。

参考文献

(1) 岡本安晴 「データ分析のための統計学入門」 株式会社おうふう 2009
(2) 河口至商 「数学ライブラリー 32 多変量解析入門Ⅰ」 森北出版 1977
(3) 杉山高一，牛沢賢二 「パソコンによる統計解析」 朝倉書店 1989
(4) 矢野恒太記念会編 「日本の100年」 国勢社 2000
(5)「データで見る県勢」 矢野恒太記念会 2002

A Study of Lectures on Educational Statistics (second report)

Tsutomu Doi

This report introduces the contents of a series of lectures given by the author for the educational statistics classes. This paper focuses on two themes: first is the “correlation between pared data” and the second is “regression analysis.” Among these themes special attention was paid to “overly strong correlation”, “spurious correlation” and “significant level in regression analysis.”

Both themes are presented by first giving numerical examples, followed by an intuitive consideration of the data, and finally the statistical theories pertinent to the data. Excel software was used during the data analysis. The author stresses the importance for the students to explain the output results using their own words.

The author believe lectures on this subject should deal with a limited number of themes rather than trying to cover too many points. This method encourages students to fulfil course requirements. Consequently their achievements in class might encourage them into farther studying in this area.