研究の世界B

2007年11月1日

Excelを用いた回帰分析

（情報フルエンシークラスと合同授業）

京都大学高等教育研究開発推進センター
第3部門　教務補佐員
持元　江津子

体験：回帰分析

回帰分析によって仮説を検証する

仮説：

エネルギー消費量はその国の豊かさによって決まるのではないか？

豊かな国ほどエネルギー消費量が多いのではないか？

その国の国民の豊かさがエネルギー消費量を説明するのではないか？

体験：回帰分析

政府統計を利用する

データの探し方：

政府統計の総合リンクサイトの統計データ・ポータルサイトを活用してみよう。

http://portal.stat.go.jp/

体験：回帰分析

環境省Web公表の環境統計集を活用

使用データ：『平成19年版　環境統計集』より，

「1.9 国・地域別　人口・面積・GDP」と「1.13 国・地域別　一時エネルギー生産量・エネルギー消費量」のExcelファイルをダウンロードし，必要なデータをExcelの新しいブックにコピペして表1を作成する。

http://www.env.go.jp/doc/toukei/index.html

表1　2002年の1人当たりのエネルギー消費量と国内総生産(GDP)

国（地域）	1人当たりGDP（米ドル）	1人当たりエネルギー消費量（石油換算：kg）
日本	30,733	3,725
インド	481	316
インドネシア a	932	456
韓国	11,570	3,504
サウジアラビア b	8,306	5,610
タイ	2,025	1,099
中国	1,009	687
トルコ	2,620	952
パキスタン	495	281
フィリピン	956	370
マレーシア	3,970	2,182
アメリカ合衆国	36,124	7,717
カナダ	23,506	7,876
メキシコ	6,300	1,332
アルゼンチン	# 2,711	1,533
ブラジル	2,576	728
アイルランド	30,649	3,718
イタリア c	20,496	3,077
英国	26,642	3,767
オーストリア	25,582	3,241
オランダ	26,072	* 5,230
ギリシャ	12,068	2,719
スイス d	38,363	3,253
スウェーデン	27,060	4,477
スペイン	16,488	2,843
チェコ	7,196	3,790
デンマーク	31,802	3,051
ドイツ	24,509	3,888
ノルウェー e	41,819	* 5,611
ハンガリー	6,377	2,519
フィンランド	25,339	5,472
フランス f	24,387	4,108
ベルギー	23,709	4,941
ポーランド	4,959	2,180
ポルトガル	11,778	2,004
ロシア	2,377	4,142
エジプト	1,204	794
ナイジェリア	2,153	154
南アフリカ g	2,374	2,771
オーストラリア h	20,469	5,732
ニュージーランド	15,341	4,451

http://www.env.go.jp/doc/toukei/index.html　参照。
練習のため元データに付随する詳細な注意書きは無視する。
1人当たりGDPはその国の国民の豊かさを測る経済指標のひとつである。

体験：回帰分析

Excelを用いて表１の散布図を描く

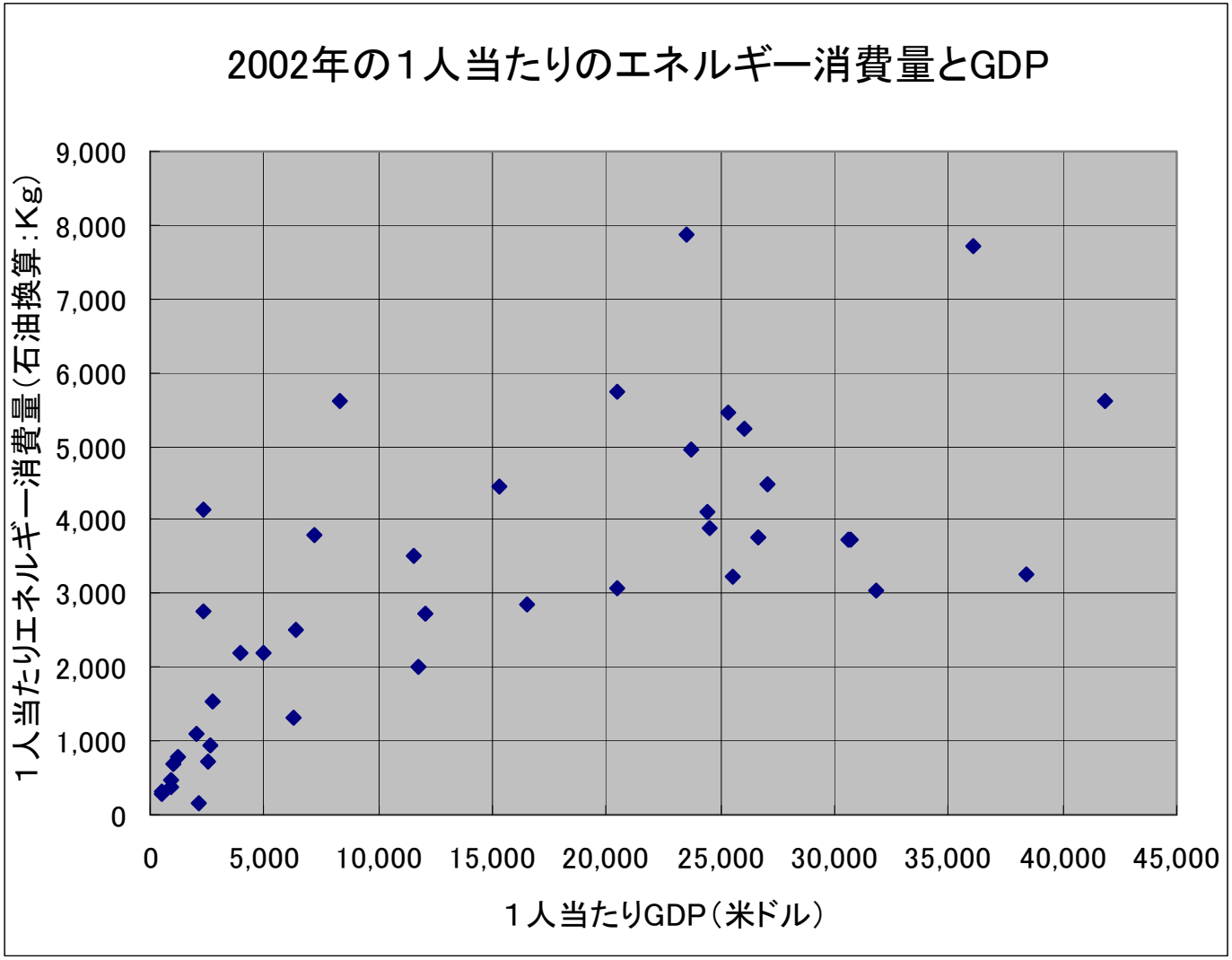
変数について：

説明変数　　１人当たりGDP

被説明変数　１人当たりエネルギー消費量

作図について，人に見せるグラフを作成する場合はメモリ数値を３～４桁までに抑える工夫をする。

体験：回帰分析


2002年の１人当たりのエネルギー消費量とGDP
１人当たりエネルギー消費量（石油換算：Kg）
9,000
8,000
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
0
5,000
10,000
15,000
20,000
25,000
30,000
35,000
40,000
45,000
１人当たりGDP（米ドル）

体験：回帰分析

散布図から集団の特徴を検討する

1．プロットされた各点の位置に注目する

　あまりにバラバラであれば回帰分析を行う前に仮説を見直さざるを得ないだろう。この散布図では…。

2．相関関係の有無について見通しをつける

　右上がりの分布から2変数間には正の相関関係がありそうと見当がつく。

3．集団が1つか複数かの見通しをつける

　データの散らばり具合に煙がたなびくような歪みがあるので，複数の集団に分けられるかもしれない。

体験：回帰分析

Excelの分析ツールを使用する

１）２変数について回帰分析を行う

２）分析結果（表２）の検討

今回は相関と決定係数に注目する

相関Rは 0.7209 であるため，まあまあの正の相関があると見てよい

（２変数に関する分析であるため「重」の文字は無視する）

体験：回帰分析

表2　表1の2変数間の回帰分析結果（概要）

回帰統計	
重相関 R	0.720931
重決定 R2	0.519742
補正 R2	0.507427
標準誤差	1398.734
観測数	41

分散分析表

	自由度	変動	分散	〟された分	有意 F
回帰	1	82574847	82574847	42.2063	1.06E-07
残差	39	76301861	1956458		
合計	40	1.59E+08			

	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	1397.47	338.8693	4.123922	0.000189	712.0426	2082.898	712.0426	2082.898
1人当たり	0.114336	0.017599	6.496637	1.06E-07	0.078738	0.149933	0.078738	0.149933

体験：回帰分析

「相関」について

ある変数の変量が増えるに連れて別の変数の変量が増えるまたは減る傾向が観察されたなら，当該の２変数間には相関関係があるという。

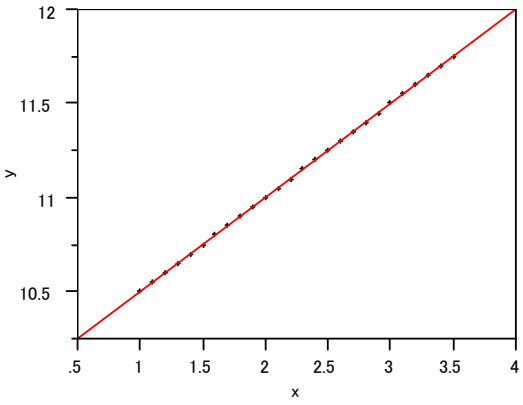
相関関係には正と負があり，右肩上がりなら正，左肩上がりなら負である。また，絶対値が「１」に近いほど相関が強く，「０」に近いほど弱い。「０」であれば相関は全くない。

相関関係が観察されたら直ちに因果関係があると思ってはいけない。偶然かも知れないし，第３の要因があるかもしれない。長期的に観察すれば，同じ傾向を示す変数がたくさんあることも少なくない。

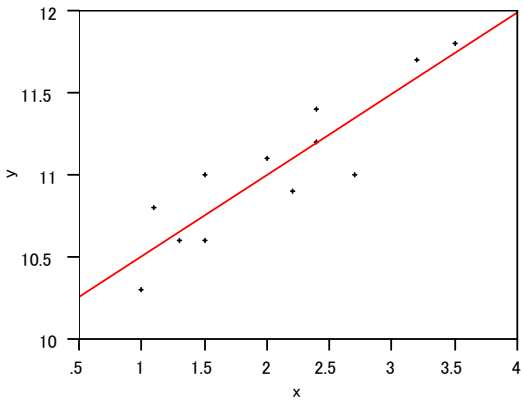
とはいえ，相関関係があれば，そこに因果関係が隠れているかもしれない。よし，もっとくわしく調べてみよう！ 相関はそういうきっかけを与えてくれるものである。

以下は，6 種類の仮想サンプルデータを基に描いた相関図に回帰直線を当てはめた図である。

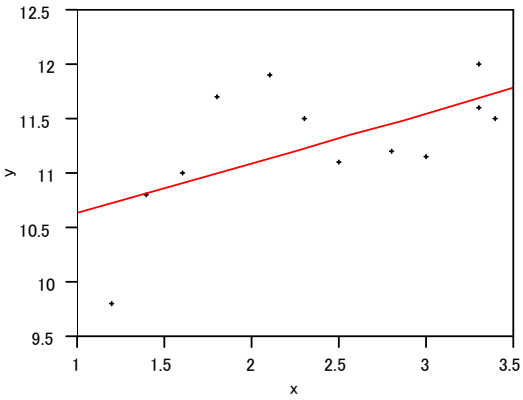
a)



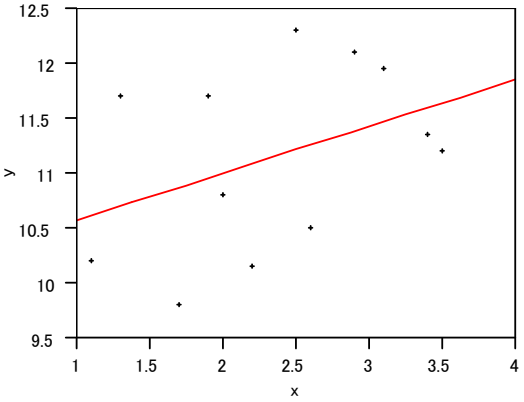
b)



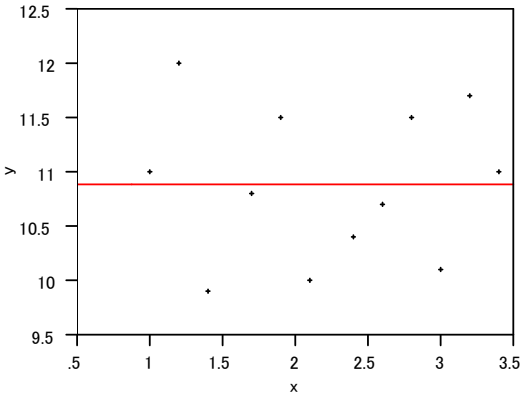
c)



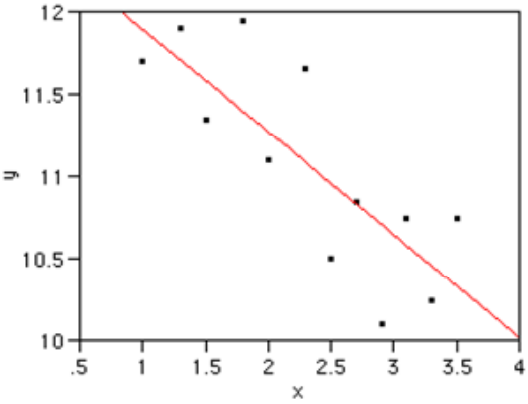
d)



e)



f)



相関係数は，図 a)から順に，1，0.9，0.6，0.4，0，−0.8 である。
相関係数の絶対値が小さいほどデータはバラバラに散らばっており，
相関係数の絶対値が 1 に近づくほどデータは回帰直線の近くに集まってくる。

体験：回帰分析

2．分析結果（表2）の検討の続き

決定係数は 0.5197 であり，いまひとつである。

決定係数：

得られた回帰直線の当てはまりのよさを示す。
説明変数が被説明変数をどれくらいよく説明しているかを示す指標。「1」に近いほどよく説明し，「0」に近いほど説明していない。寄与率とも言う。

体験：回帰分析

集団の検討

今行ったばかりの回帰分析は表１のデータが１集団を表す前提である。これで本当によいだろうか？

先に描いた散布図は，複数の集団の存在を示唆しているようにも見えた。

複数の集団の有無について検討してみよう。

複数の集団に切り分けるとすれば，どこに境界線をおけばよいのだろうか？

体験：回帰分析

集団の検討（続き）

切り分けの方法：

（1） 何らかの概念や常識に沿って切り分ける
（社会科学的な研究の場合に多く，境界線を自在に移動させることが難しい。）

（2） データから都合のよい境界を探索する
（最も都合のよい境界線がみつかるまで，境界値を自在に動かす。とりわけ工学的な分野では常套である。）

体験：回帰分析

集団の検討（続き２）

　ここでは社会科学的な仮説を立てている。よって，（１）の切り分け方法を採用する。

　豊かな国とそうでない国を切り分けることとする。OECDなどの国際機関の見解を集約した１人当たりGDP１万米ドルを境界値とする。

　ちょうど韓国とサウジアラビアの間が境界となる。

体験：回帰分析

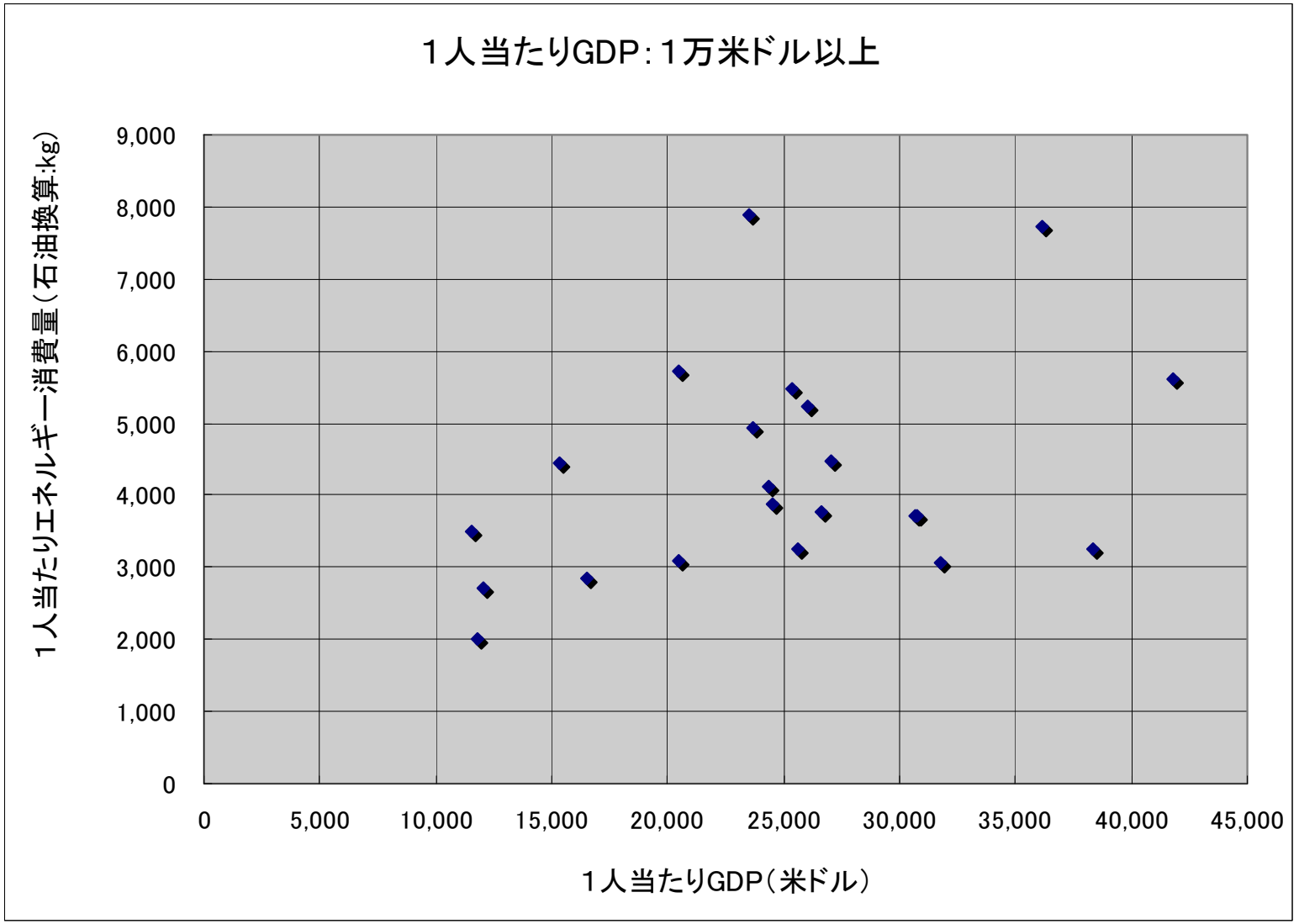
２集団の検討

先ほど見つけた境界で切り分けて，１人当たりGDPが1万米ドル以上の豊かな国々のグループと，1万米ドル未満の相対的に貧しい国々のグループを設定した。

それぞれについて検討する。

具体的には，それぞれのグループに対応する散布図を描き，Excelによる回帰分析を行う。

体験：回帰分析


１人当たりGDP：１万米ドル以上
１人当たりエネルギー消費量（石油換算:kg）
9,000
8,000
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
0
5,000
10,000
15,000
20,000
25,000
30,000
35,000
40,000
45,000
１人当たりGDP（米ドル）

体験：回帰分析

表3　1人当たりGDP1万米ドル以上：回帰分析結果（概要）

回帰統計	
重相関 R	0.369048
重決定 R2	0.136196
補正 R2	0.093006
標準誤差	1438.274
観測数	22

分散分析表

	自由度	変動	分散	観された分散	有意 F
回帰	1	6523231	6523231	3.153402	0.090988
残差	20	41372653	2068633		
合計	21	47895885			

	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	2623.525	987.882	2.655707	0.015177	562.839	4684.21	562.839	4684.21
1人当たり	0.067377	0.037942	1.775782	0.090988	−0.01177	0.146524	−0.01177	0.146524

体験：回帰分析

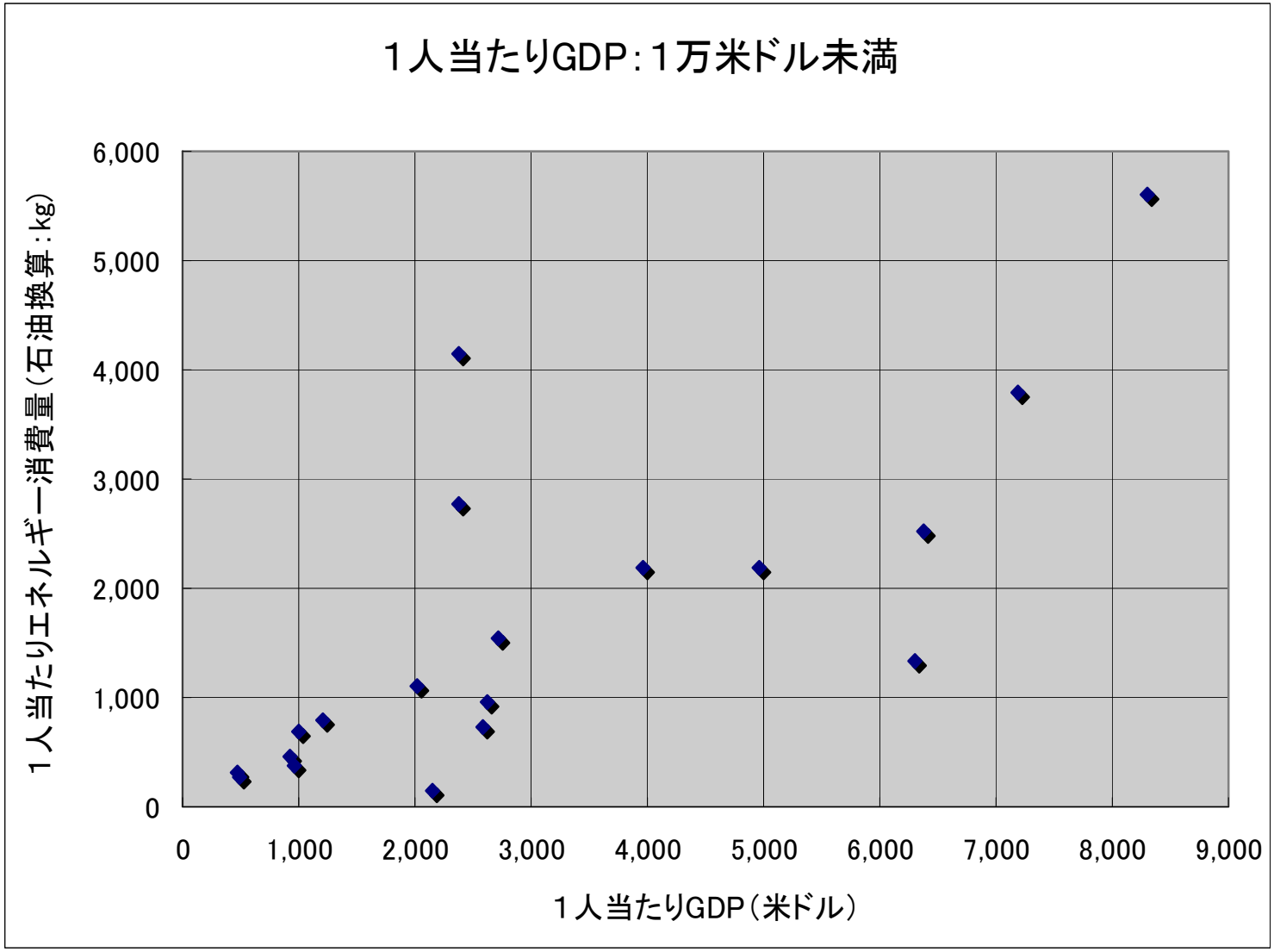
豊かな国々の場合

相関係数も決定係数も切り分ける前に比べて悪くなっている。

つまり，このグループに関して，回帰分析の適用がふさわしくない。

まったく別の分析方法を使うべきかもしれない。

体験：回帰分析


１人当たりGDP：１万米ドル未満
１人当たりエネルギー消費量（石油換算：kg）
6,000
5,000
4,000
3,000
2,000
1,000
0
0
1,000
2,000
3,000
4,000
5,000
6,000
7,000
8,000
9,000
１人当たりGDP（米ドル）

体験：回帰分析

表4　1人当たりGDP1万米ドル未満：回帰分析結果（概要）

回帰統計	
重相関 R	0.746809
重決定 R2	0.557723
補正 R2	0.531707
標準誤差	1037.858
観測数	19

分散分析表

	自由度	変動	分散	観された分散	有意 F
回帰	1	23091339	23091339	21.43746	0.000239
残差	17	18311537	1077149		
合計	18	41402876			

	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	213.0989	396.0997	0.537993	0.597555	-622.598	1048.796	-622.598	1048.796
1人当たり	0.471809	0.101901	4.63006	0.000239	0.256816	0.686803	0.256816	0.686803

体験：回帰分析

相対的に貧しい国々の場合

相関係数も決定係数も改善した。

ここで得られた回帰直線式は，yを１人当たりエネルギー消費量，xを１人当たりGDPとして，

y＝0.4718x＋213.0989 である。

この回帰モデルの当てはまりのよさの度合いを表すのが決定係数なのだが，改善しているとはいえ，まだまだ不十分な水準である。

体験：回帰分析

次に何をすべきか？

少なくとも1万米ドル未満のグループでは，初めに立てた仮説がそこそこ当たっているように見える。今すぐこの仮説を捨ててしまうのは惜しい気がする。

・国際機関の公表するデータなどを調べて，もっとたくさんの国・地域について調べてデータセット数を増やす。
・新しく変数を加えて多重回帰分析を行ってみる 。
・仮説の一部を変更してきりわけの境界値を移動させる。（ex.近年目覚しく経済発展を遂げてきている中国などを目印にする） etc.

おわりに

Excelによる回帰分析は，調査研究の初期段階において，研究の方向を探るのに大変便利である。大いに活用されたい。

しかし，研究を本格的に進め，学術的な研究論文として発表する際には，もっと信頼性の高いSPSSやSASといった統計ソフトを使うべきである。

また，Excelによる回帰分析に関する解説本やWebサイトも数多く存在するので，それらも参照しながら回帰分析について学ぼう。