データ分析基礎

回帰分析 演習

京都大学 国際高等教育院 關戸啓人

■使用するデータについて：

東京の日平均気温の月平均値：
http://www.data.jma.go.jp/obd/stats/etrn/view/monthly_s3.php?prec_no=44&block_no=47662

京都の日平均気温の月平均値：
http://www.data.jma.go.jp/obd/stats/etrn/view/monthly_s3.php?prec_no=61&block_no=47759

アサヒグループホールディングスの月次販売情報
https://www.asahigroup-holdings.com/ir/financial_data/monthly_data.html

■目的：
気温とビールの売り上げは関係があるといわれている.
実際にデータを分析することでその関係を見てみよう.
そのような関係を調べることによって，例えば，気象庁の長期予報などと組み合わせることによって，ビール製造会社がビールの売り上げを予想でき，どの程度製造すべきかの目安となったり，飲食店がビールをどれぐらい仕入れるかの目安となったりするかもしれない.

■使用する手法・キーワードなど：
回帰分析（単回帰分析，重回帰分析）
多重共線性
ダミー変数
Ridge 回帰，Lasso 回帰
モデル選択（AIC，クロスバリデーション）

■注意：
以下の手順には，学習のために反面教師としてふるまっている部分もあり，データの分析の手順として必ずしも良いものではない.

■手順：

□　データのダウンロードと整形

アサヒグループホールディングスのホームページから，2011 年 1 月から 2017 年 4 月までのスーパードライの月次販売情報をダウンロード（コピーアンドペースト）する．
また，同様に，気象庁のページから，同じ期間の東京の日平均気温の月平均値をダウンロード（コピーアンドペースト）する．
ダウンロードしたデータを整形し，csv ファイルを作成する．
作成した csv ファイルのサンプルは以下のとおりである．
http://ds.k.kyoto-u.ac.jp/e-learning_files/data_analysis_basic/jma_001.csv

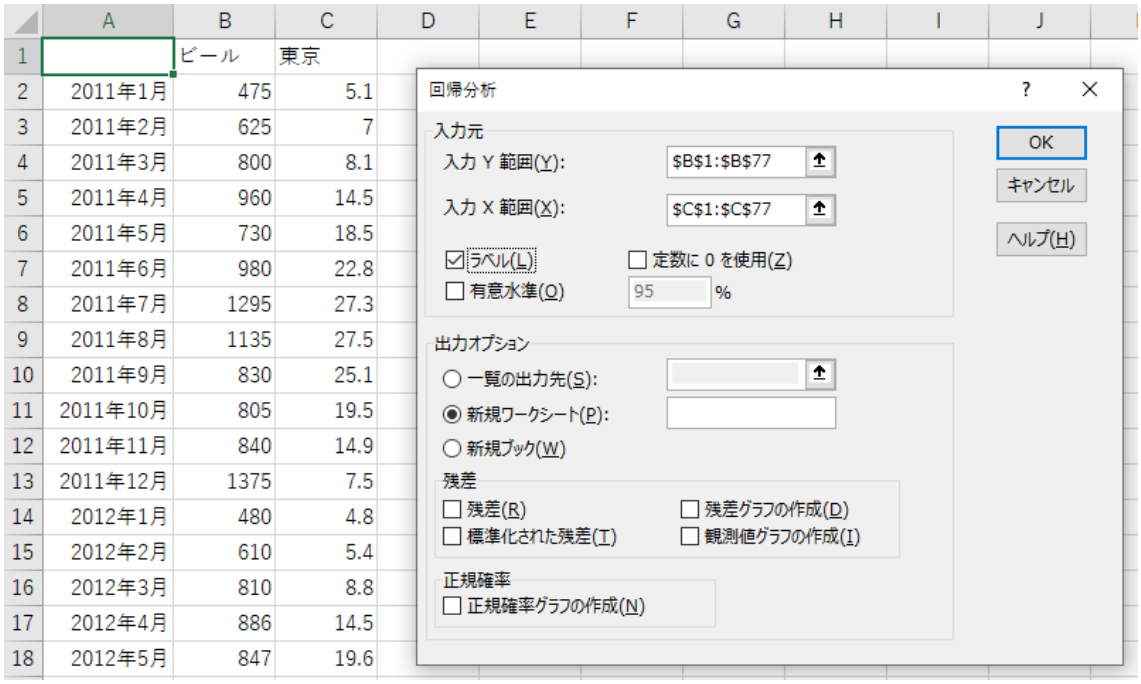
※必ずしもダウンロードするデータの期間は 2011 年 1 月から 2017 年 4 月でなくても良いが，多重共線性の説明の際に，結果が異なる可能性がある．

□　単回帰分析の実行

作成した csv ファイルを Excel で開き，Excel を用いて回帰分析を行う．
Excel を用いて回帰分析する方法はいくつかあるが，ここでは分析ツールのアドインを使用する．

まずは，以下の手順により，分析ツールのアドインを有効化する．
ファイル → オプション → アドイン → 設定 → 分析ツールにチェックを入れて OK を押す

分析ツールのアドインが有効化されると，リボンのデータのタブにデータ分析が表示される．
データ分析をクリックし，回帰分析を選び，OK と押す．
入力 Y 範囲，入力 X 範囲を適切に記入し，OK を押すことで，回帰分析を行う．
上の URL からダウンロードした csv ファイルを利用する場合は，以下のように記入・変更すれば良い：
入力 Y 範囲：$B$1:$B$77　　（$ はなくても良い）
入力 X 範囲：$C$1:$C$77　　（$ はなくても良い）
「ラベル」にチェックをいれる



	A	B	C
1		ビール	東京
2	2011年1月	475	5.1
3	2011年2月	625	7
4	2011年3月	800	8.1
5	2011年4月	960	14.5
6	2011年5月	730	18.5
7	2011年6月	980	22.8
8	2011年7月	1295	27.3
9	2011年8月	1135	27.5
10	2011年9月	830	25.1
11	2011年10月	805	19.5
12	2011年11月	840	14.9
13	2011年12月	1375	7.5
14	2012年1月	480	4.8
15	2012年2月	610	5.4
16	2012年3月	810	8.8
17	2012年4月	886	14.5
18	2012年5月	847	19.6

□　単回帰分析の結果を読み取る

単回帰分析を行った結果は以下のような感じになるはずである.

	A	B	C	D	E	F	G	H	I
1	概要								
2									
3	回帰統計								
4	重相関 R	0.431346							
5	重決定 R2	0.186059							
6	補正 R2	0.17506							
7	標準誤差	212.299							
8	観測数	76							
9									
10	分散分析表								
11		自由度	変動	分散	測された分散	有意 F			
12	回帰	1	762406.1	762406.1	16.91572	0.0001			
13	残差	74	3335243	45070.86					
14	合計	75	4097650						
15									
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
17	切片	656.9307	56.52093	11.62279	2.33E-18	544.3103	769.5511	544.3103	769.5511
18	東京	12.99158	3.15876	4.112873	0.0001	6.697612	19.28555	6.697612	19.28555

上図の場合で結果を少し見ていくと：

東京の係数が 13 程度であるので，1 度気温が上がるごとに売り上げが 13（万箱）程度増えるであろうということが読み取れる．
また，P 値（0.0001）や 95%信頼区間（[下限 95%，上限 95%] = [6.697612, 19.28555]）を見てやると，確かに東京の気温とビールの売り上げの間には関係がありそうである，ということや，東京の気温が変化したときに，少なくても（例えば，確率 95%以上で）どれぐらいビールの売り上げが変化するだろうか，というのに役に立ちそうな情報も得られそうである．

ただし，これらは，「ビールの売り上げ」は「東京の気温」で説明されるという線形回帰モデルを前提とした分析結果であり，そこには色々な仮定が置かれていることに注意する．
また，必ずしも，回帰分析は因果的な結果を意味しているわけでもないことに注意する．

ところで，重決定 R2 の値を見てやると 0.186 程度とかなり低い値になっている．
これはビールの売り上げの変動（合計の変動=4097650）のうち，東京の気温の変化によって説明できる変動（回帰の変動=762406.1）の割合を意味しており，ビールの売り上げについて，東京の気温で説明しようとしてもあまり説明できていないことを意味している．
そこで，次に，もっとビールの売り上げをよく説明するために，説明変数を増やしてみることを考えてみよう．

□　重回帰分析のための準備
東京の気温のデータのみでは，ビールの売り上げをうまく説明できなかったので，京都の気温のデータも使用し，重回帰分析してみよう．
東京の気温のデータをダウンロードしたのと同様に，気象庁のページから，京都の日平均気温の月平均値をダウンロード（コピーアンドペースト）する．
ダウンロードしたデータを整形し，csv ファイルを作成する．
作成した csv ファイルのサンプルは以下のとおりである．
http://ds.k.kyoto-u.ac.jp/e-learning_files/data_analysis_basic/jma_002.csv

□　重回帰分析の実行
重回帰分析を行うには，分析ツールの回帰分析を使用するときに，入力 X 範囲で複数の列を指定すれば良い．
上の URL からダウンロードした csv ファイルを利用する場合は，以下のように記入・変更すれば良い：
入力 Y 範囲：$B$1:$B$77　　（$ はなくても良い）
入力 X 範囲：$C$1:$D$77　　（$ はなくても良い）
「ラベル」にチェックをいれる

□　重回帰分析の結果を読み取る

重回帰分析の結果は以下のようになるはずである．

	A	B	C	D	E	F	G	H	I
1	概要								
2									
3	回帰統計								
4	重相関 R	0.431415							
5	重決定 R2	0.186119							
6	補正 R2	0.163821							
7	標準誤差	213.7403							
8	観測数	76							
9									
10	分散分析表								
11		自由度	変動	分散	測された分散	有意 F			
12	回帰	2	762649.4	381324.7	8.346836	0.000544			
13	残差	73	3335000	45684.94					
14	合計	75	4097650						
15									
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
17	切片	653.1112	77.32065	8.446788	2.04E-12	499.0113	807.211	499.0113	807.211
18	東京	15.34582	32.42195	0.473316	0.6374	-49.271	79.96268	-49.271	79.96268
19	京都	-2.15779	29.57313	-0.07296	0.942034	-61.097	56.78137	-61.097	56.78137

まず，目的であった，重決定 R2 の値を確認すると，ほぼ同じ値であるので全く改善されていない．

（それどころか，補正 R2 で見ると，悪くなっている）

また，京都の係数を見ると，京都の気温が 1 度上がるとビールの売り上げが 2（万箱）程度減るという，一見不可解な結果になっている．

実際のところは，東京，京都の信頼区間などを見てやると，推定結果が不安定になっているように見える．

これは東京と京都の気温の間は高い正の相関があり，多重共線性という現象が起こっており，東京と京都の気温の両方を説明変数として使用するのはあまり良くないと思われる．

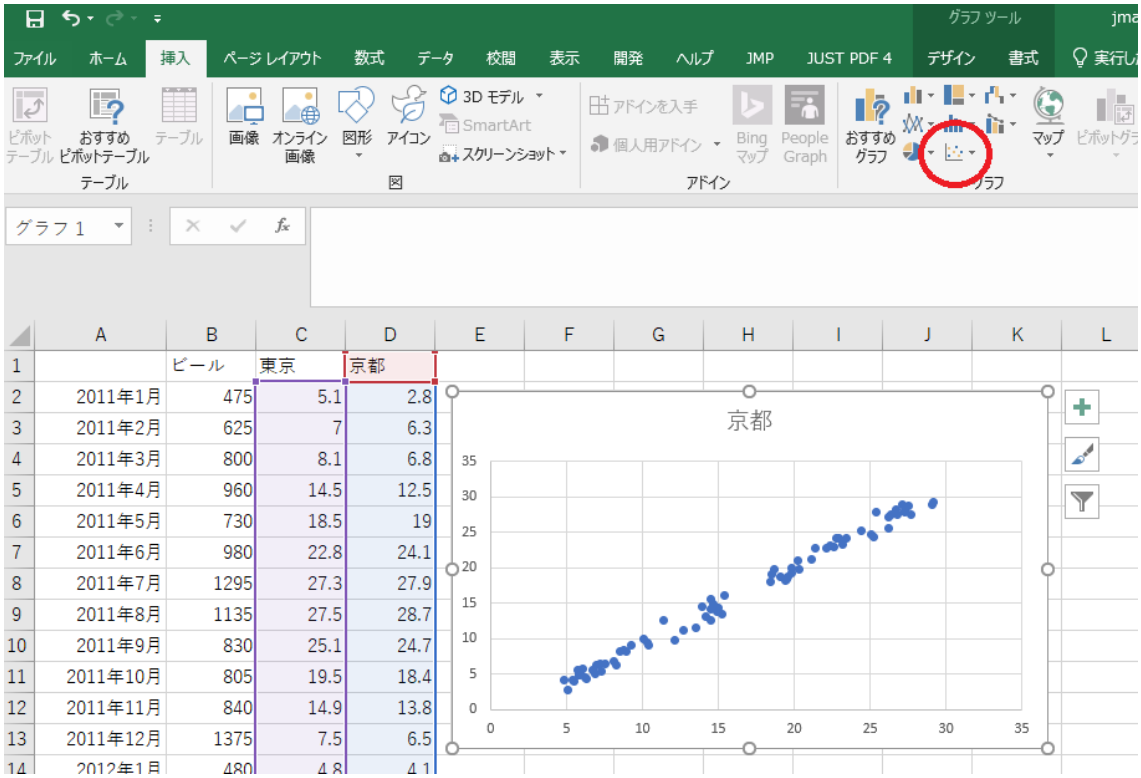
□　散布図を確認

Excel ではいくつかのデータの可視化を手軽に行うことができる．

例えば，東京の気温と京都の気温の 2 列を選択して，

挿入→グラフ→散布図

と操作することで，東京の気温と京都の気温の散布図を描くことができる．



	A	B	C	D
1		ビール	東京	京都
2	2011年1月	475	5.1	2.8
3	2011年2月	625	7	6.3
4	2011年3月	800	8.1	6.8
5	2011年4月	960	14.5	12.5
6	2011年5月	730	18.5	19
7	2011年6月	980	22.8	24.1
8	2011年7月	1295	27.3	27.9
9	2011年8月	1135	27.5	28.7
10	2011年9月	830	25.1	24.7
11	2011年10月	805	19.5	18.4
12	2011年11月	840	14.9	13.8
13	2011年12月	1375	7.5	6.5
14	2012年1月	480	4.8	4.1

この散布図を見ると，東京の気温と京都の気温は高い正の相関があることが視覚的に確かめられる．

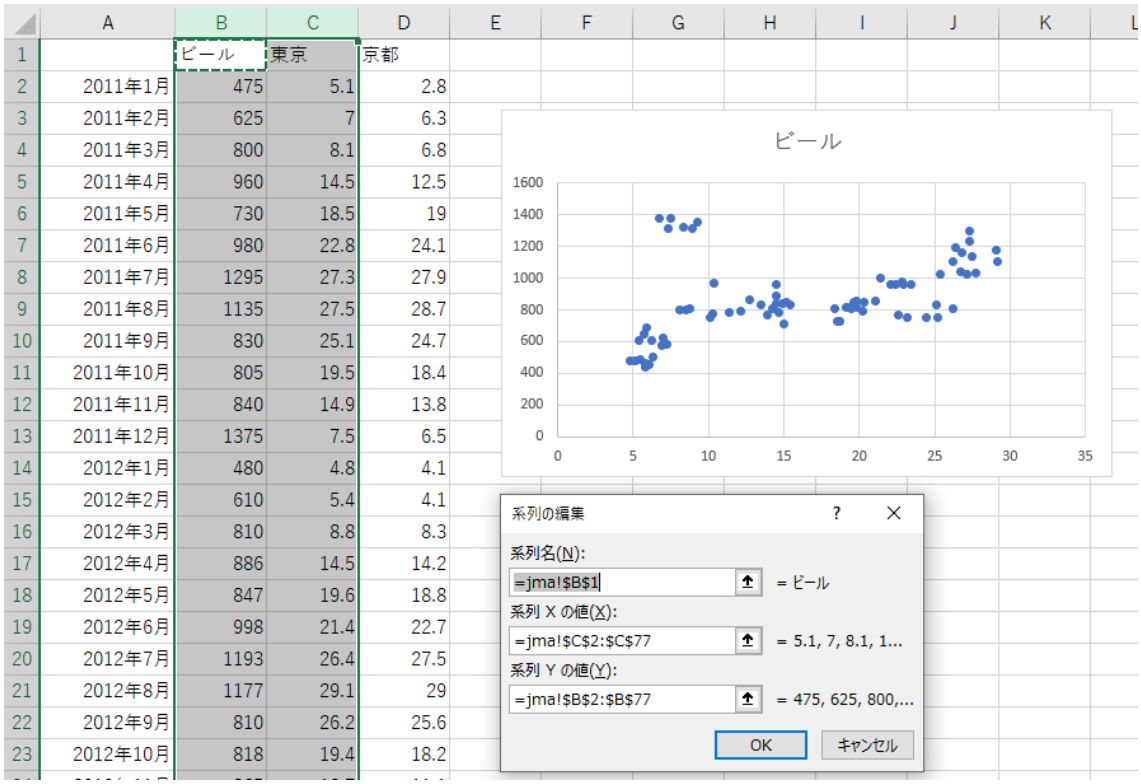
では，ビールの売り上げと東京の気温の散布図を描くことで，なぜうまく説明できないのかを考えてみよう．

回帰分析的にはビールの売り上げを y 軸（縦軸），東京の気温を x 軸（横軸）として散布図を描いてみたいが，Excel の散布図では，左の列が x 軸，右の列が y 軸に取られるので，少し工夫が必要である．

先ほど作成した散布図を右クリックし，

データの選択→編集

から，直接 x 軸，y 軸を指定してあげても良いし，東京の売り上げが左に，ビールの売り上げが右の列になるように，適当にコピーアンドペーストでデータを移動させても良い．



	A	B	C	D
1		ビール	東京	京都
2	2011年1月	475	5.1	2.8
3	2011年2月	625	7	6.3
4	2011年3月	800	8.1	6.8
5	2011年4月	960	14.5	12.5
6	2011年5月	730	18.5	19
7	2011年6月	980	22.8	24.1
8	2011年7月	1295	27.3	27.9
9	2011年8月	1135	27.5	28.7
10	2011年9月	830	25.1	24.7
11	2011年10月	805	19.5	18.4
12	2011年11月	840	14.9	13.8
13	2011年12月	1375	7.5	6.5
14	2012年1月	480	4.8	4.1
15	2012年2月	610	5.4	4.1
16	2012年3月	810	8.8	8.3
17	2012年4月	886	14.5	14.2
18	2012年5月	847	19.6	18.8
19	2012年6月	998	21.4	22.7
20	2012年7月	1193	26.4	27.5
21	2012年8月	1177	29.1	29
22	2012年9月	810	26.2	25.6
23	2012年10月	818	19.4	18.2

さて，今回はこのタイミングでデータの可視化を行ったが，本来は最初に行うことをおすすめする．
まずはデータがどのようなものなのかを確かめるのも重要であるし，常識的に考えにくいデータ（例えば気温が 100℃を超えているなど）がないことを確かめるのも重要である．

□　今までの分析のおける問題点とダミー変数の利用による解決
今，扱っている，ビールの売り上げと東京の気温のデータも可視化してみて，回帰分析を行うことが妥当かどうかを見てみたい．
左上の 6 つのデータを除けば，だいたい気温が高くなるほどビールの売り上げが単調に増えているように見える．では，左上のデータは何かを確認してみると，全て 12 月に該当するデータであることがわかる．
確かに，12 月では，気温はとても低いのにもかかわらず，ビールの売り上げは非常に多い．その理由としては，忘年会，お歳暮，新年会の準備などが考えられる．
そこで，1 つの仮説として，12 月のデータがその他の月と違う振る舞いをすることで回帰分析があまりうまく行かなかったのではないか，と考えてみよう．

では，解決策はどのようなものが考えられるだろうか．
例えば，12 月が通常と違う振る舞いをするのだから，12 月のデータを削除してから，回帰

分析を行う，というのも 1 つの手である．
ここでは，説明変数に，12 月かどうかを表すダミー変数を加えることで解決を試みよう．
つまり，Excel の表に 12 月のときに 1，その他のときに 0 という列を加えてやれば良い．
この列を加え，京都の列を削除したものが，
http://ds.k.kyoto-u.ac.jp/e-learning_files/data_analysis_basic/jma_003.csv
である．

これで回帰分析を行うと以下のようになるはずである．

	A	B	C	D	E	F	G	H	I
1	概要								
2									
3	回帰統計								
4	重相関 R	0.883829							
5	重決定 R2	0.781153							
6	補正 R2	0.775157							
7	標準誤差	110.8348							
8	観測数	76							
9									
10	分散分析表								
11		自由度	変動	分散	測された分散	有意 F			
12	回帰	2	3200891	1600446	130.2832	8.22E-25			
13	残差	73	896758.4	12284.36					
14	合計	75	4097650						
15									
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
17	切片	479.715	32.07686	14.95518	6.34E-24	415.7859	543.6441	415.7859	543.6441
18	東京	20.55118	1.734185	11.85062	1.12E-18	17.09495	24.0074	17.09495	24.0074
19	12月	698.5422	49.58027	14.08912	1.63E-22	599.7289	797.3556	599.7289	797.3556

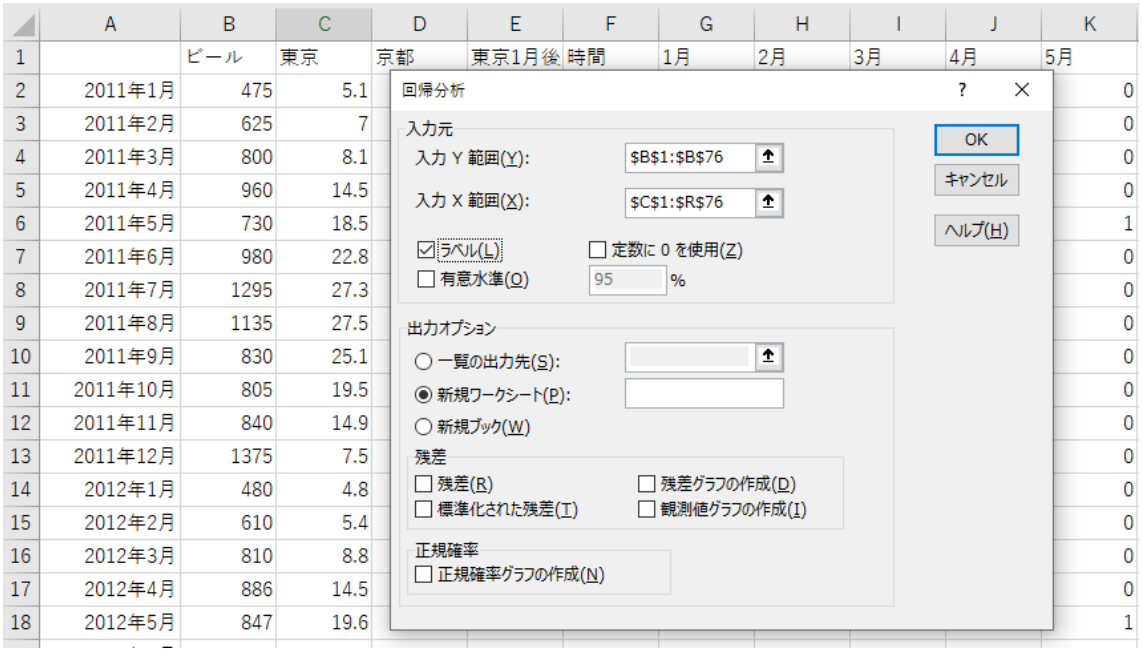
12 月を表すダミー変数をいれることで，重決定 R2 の値が飛躍的に改善したことが見て取れる．

□　更なる変数の追加の検討

さて，これで十分だろうか．12 月だけ特別視するのは違和感がないだろうか．例えば 3 月や 4 月には送別会や歓迎会などが数多く行われるだろう．また，8 月は暑いからビールを飲んでいるわけではなく，8 月はビールを飲む時期だ，と思い込んでビールを飲んでいる場合は，気温ではなく何月かを表すダミー変数を使用した方がうまくビールの売り上げを説明できるかもしれない．

そこで，色々と説明変数を加えてみよう．
1 月～12 月までを表すダミー変数 12 種と，単純に時間を表す変数（近年になるほどビール離れが進む，などを考慮している），次の月の東京の気温（次の月の準備を考慮している）を加えてみた csv ファイルを作成してみたのが以下のものである（次の月の値を使用するデータを加えた関係で，観測数は 1 減っている）．
http://ds.k.kyoto-u.ac.jp/e-learning_files/data_analysis_basic/jma_004.csv
他にも，例えば，ビールを飲むことができる 20 歳以上の人口だとか，該当する月の平日や休日・祝日の日数など（これらは比較的容易に未来の数値を予測できる）を加えることも考えられるだろう．また，散布図の結果から，ビールの売り上げの関係は，気温の 3 次関数ではないか，などと思った場合は，気温の 2 乗や 3 乗を説明変数として加えることも考えても良いだろう．
ただし，Excel で回帰分析を行う場合は，説明変数を 16 個までしか指定することができない．その以上の説明変数を使いたい場合は，他のソフトウェア等を使った方が良いかもしれない．

さて，上の csv ファイルで回帰分析を行うには，



の陽にすればよく，その結果は以下の通りになる．

	A	B	C	D	E	F	G	H	I
1	概要								
2									
3	回帰統計								
4	重相関 R	0.984428							
5	重決定 R2	0.969098							
6	補正 R2	0.944292							
7	標準誤差	46.2881							
8	観測数	75							
9									
10	分散分析表								
11		自由度	変動	分散	測された分散	有意 F			
12	回帰	16	3964305	247769	123.3494	1.12E-38			
13	残差	59	126412.7	2142.588					
14	合計	75	4090717						
15									
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
17	切片	684.3485	105.1918	6.505719	1.85E-08	473.8601	894.8368	473.8601	894.8368
18	東京	17.56614	9.091125	1.93223	0.058137	-0.62515	35.75744	-0.62515	35.75744
19	京都	-9.73623	9.133442	-1.066	0.290768	-28.0122	8.539748	-28.0122	8.539748
20	東京1月後	5.127993	6.523588	0.786069	0.434973	-7.92568	18.18166	-7.92568	18.18166
21	時間	-0.87056	0.296181	-2.93927	0.00469	-1.46321	-0.2779	-1.46321	-0.2779
22	1月	-270.33	61.10781	-4.42383	4.24E-05	-392.607	-148.054	-392.607	-148.054
23	2月	-141.868	43.39417	-3.2693	0.001802	-228.7	-55.037	-228.7	-55.037
24	3月	0	0	65535	#NUM!	0	0	0	0
25	4月	-31.3622	50.58889	-0.61994	#NUM!	-132.59	69.86595	-132.59	69.86595
26	5月	-103.683	82.74835	-1.25299	0.215153	-269.262	61.89637	-269.262	61.89637
27	6月	13.65431	110.5861	0.123472	0.902153	-207.628	234.9365	-207.628	234.9365
28	7月	177.2383	135.3375	1.309603	0.195407	-93.5713	448.048	-93.5713	448.048
29	8月	97.66896	134.6737	0.725226	0.47118	-171.813	367.1505	-171.813	367.1505
30	9月	-163.846	105.6793	-1.55041	0.126391	-375.31	47.61743	-375.31	47.61743
31	10月	-72.7804	75.65567	-0.962	0.339978	-224.167	78.60622	-224.167	78.60622
32	11月	15.48542	63.00444	0.245783	0.806703	-110.586	141.557	-110.586	141.557
33	12月	591.406	62.96016	9.393337	2.54E-13	465.423	717.389	465.423	717.389

さて，3 月 4 月のダミー変数のあたりの結果が何やら変なことになっているが，取り敢えず他の部分を見ていく．

東京・京都は相変わらず多重共線性のため信用できない値ではあるが，係数の和が 17.5+(-9.73)が正であるので，やっぱり時期的な問題だけではなく，（同じ月でも）気温によってビールの売り上げは変わりそうだなあ…，という結果が見て取れる．

重決定 R2 の値はかなり 1 に近づき（補正 R2 でもかなり高い値になっており），ビールの

売り上げの変動のうち，かなりの割合が現在の説明変数で説明されていることがわかる．

ただし，重決定 R2 が高いというのは，今あるデータ（標本，サンプル）に対してうまく説明されているということであり，全体的なデータ（母集団），例えば将来のデータについてうまく説明できることを必ずしも意味しない．一般的には，説明変数を増やすと過剰適合（過学習）などのため，母集団をうまく説明できない場合がある．また，多重共線性の問題もあるし，説明変数をうまく選ばないといけない，ということになる．

さて，この流れで，3 月 4 月のダミー変数あたりの結果が変なことになっていることについて説明しよう．これは，多重共線性の極端な場合が起こっていて，推定不可能になっている．
（今あるデータについて,）全てのデータは 1 月から 12 月までのどれか 1 つの属すため，例えば，1 月～12 月までのダミー変数の係数を全て 1 増やし，切片を 1 減らしても，残差は不変であり，最小二乗推定量が一意に定まらない．
これを解決するのは，どこかの月を基準に取り，その月からの差，という意味合いを持たせると，その基準の月に対応するダミー変数を削除できる．
しかし，「説明変数は少ない方が良い」という状況では，どの月を基準に取るのが良いかは簡単な話ではないし，そもそも〇月と△月は同一視した方が良いだろうなど，色々なパターンが考えられる．
そうなると，どのような「モデル」が良いのかを真面目に考えなければならなくなる．

□ Ridge 回帰と Lasso 回帰の実行
※ここからは，正直，Excel で行うより R や Python などで行った方が良い内容です．

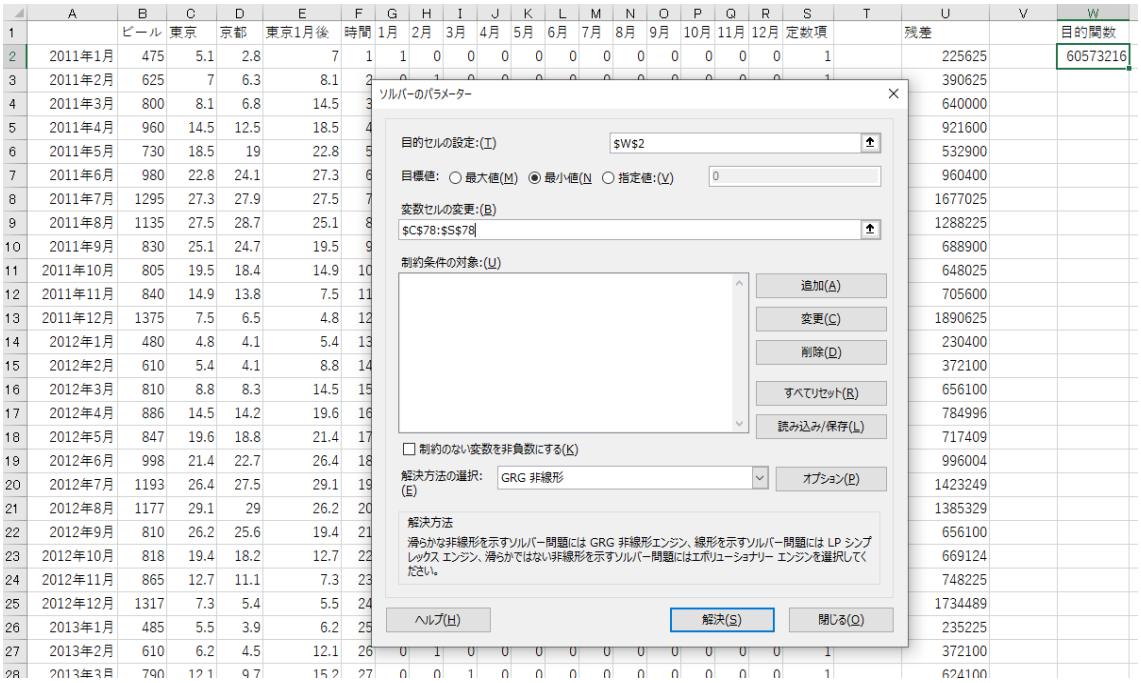
では，どの「モデル」が良いかを考える…，とする前に，適当に説明変数をがむしゃらに増やしてもある程度うまく行くという方法を先に紹介しよう．
それは，正則化項をつけることで，正則化項の入れ方で，例えば Ridge 回帰と呼ばれるものや Lasso 回帰と呼ばれるものがある（スライド参照）．

そのような正則化項をつけた回帰を行う専用の方法を Excel は提供していないが，Excel のアドインの中には一般的に最適化問題を解く「ソルバー」があり，それを利用することで行うことができる．
アドインのソルバーを有効化するには，分析ツールと同様に以下の手順で行う：
ファイル → オプション → アドイン → 設定 → ソルバーにチェックを入れて OK を押す

このソルバーは，いくつかのセルを（ある条件を満たす中で）自由に値を変化させたときに，あるセルの値を最小化・最大化・とある値にできるだけ近づける，ということをやってくれる．
これを利用して，Ridge 回帰や，Lasso 回帰の目的関数の値をどこかのセルに計算しておいて，そのセルの値を最小化することで回帰分析を行うことができる．

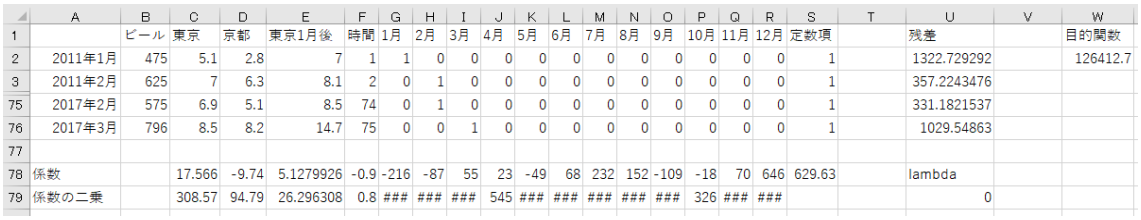
例えば，Ridge 回帰を見据えて，成型したものが
http://ds.k.kyoto-u.ac.jp/e-learning_files/data_analysis_basic/jma_005.xlsx
である．

このファイルとソルバーアドインを利用して，



のように最適化問題を解いてやることで，Ridge 回帰を行うことができる．

lambda=0 の場合は，最小二乗法と一致し，その結果はうまく行けば，



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1		ビール	東京	京都	東京1月後	時間	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月	定数項		残差		目的関数
2	2011年1月	475	5.1	2.8	7	1	1	0	0	0	0	0	0	0	0	0	0	0	1		1322.729292		126412.7
3	2011年2月	625	7	6.3	8.1	2	0	1	0	0	0	0	0	0	0	0	0	0	1		357.2243476		
75	2017年2月	575	6.9	5.1	8.5	74	0	1	0	0	0	0	0	0	0	0	0	0	1		331.1821537		
76	2017年3月	796	8.5	8.2	14.7	75	0	0	1	0	0	0	0	0	0	0	0	0	1		1029.54863		
77																							
78	係数		17.566	-9.74	5.1279926	-0.9	-216	-87	55	23	-49	68	232	152	-109	-18	70	646	629.63		lambda		
79	係数の二乗		308.57	94.79	26.296308	0.8	###	###	###	545	###	###	###	###	###	326	###	###			0		

のようになり，目的関数の値が，回帰分析した際の残差の二乗和（残差の変動）と一致していることが確認できる．
ただし，一般的には，汎用的に最適化問題を解くのは難しく，値が一致しない場合も多い．

その場合はソルバーのパラメータを変えながら何回か試すと最適解を求めることができることがある（少なくても，この問題の場合は解けるはず）.
lambda を変化させたり，Lasso 回帰を試していたりして欲しい.
例えば，lambda を大きくしすぎると，答えが「鈍る」ことや，Lasso 回帰の場合はスパース性が出てくることなどを確認する.

一般的には最適化問題を解くのは簡単ではなく，例えば Lasso 回帰を行うにはどのように最適化問題を解けば良いか，など個別な状況に特化した方法が数多く提案され，実際にそのような解法が使える場合は，その方が良いことを注意しておく.

□ VBA の利用とモデル選択の実行例
さて，正則化項を入れるのではなく，モデル選択を行うためには，一般的には試行錯誤が必要になる.
実際に，回帰分析を何回も行い，どのモデルが良いかを調べるのは大変であるので，プログラミングの力を利用するのが良いと思われる.
Excel では，本来はマクロ機能であるのだが，VBA（Visual Basic for Applications）を用いてプログラミングを行うことができる.

VBA でプログラミングをするためには，以下の手順で，リボンに開発タブを表示させる必要がある.
ファイル→オプション→リボンのユーザー設定→開発にチェックを入れる

また，VBA から回帰分析を行うには，分析ツール – VBA のアドインを有化する必要があるので，以下の手順で有効化しておく.
ファイル → オプション → アドイン → 設定 → 分析ツール – VBA にチェックを入れて OK を押す

後は，開発タブから Visual Basic を選択し，
ユーザーフォームの挿入→標準モジュール
としてできたらウインドウにプログラムを書いていけば良い.

例えば，回帰分析を行うサブルーチンは（適当に書いたものであるが）

```
Sub lm(n, y, x1, x2, out_r, out_c)
    Range(Cells(out_r, out_c), Cells(out_r + 30 + x2 - x1, out_c + 9)) = ""
    Application.Run "ATPVBAEN.XLAM!Regress", Range(Cells(1, y), Cells(n + 1, y)),
Range(Cells(1, x1), Cells(n + 1, x2)), False, True, 95, Cells(out_r, out_c), False, False,
```

```
False, False, , False
End Sub
```

となり，引数の意味は：

・n は観測数

・y は被説明変数のデータが入っている列

・x1 列から x2 列までに説明変数のデータが入っている

・出力結果は，out_r 行 our_c 列のセルを左上のセルとなるように出力する

であり，

```
Sub hoge()
    lm 75, 2, 3, 18, 1, 20
End Sub
```

とサブルーチンを作って実行することができる.

他にも，AIC を計算したり，簡単なクロスバリデーションのようなことを行うサブルーチンを作ってみると，例えば，

```
Sub predict_check(n, m, y, x1, x2, out_r, out_c)
    lm n, y, x1, x2, out_r, out_c

    er = 0
    For i = 1 To m
        correct = Cells(1 + n + i, y)
        predict = Cells(out_r + 16, out_c + 1) ' 切片
        For j = x1 To x2
            predict = predict + Cells(out_r + 17 + j - x1, out_c + 1) * Cells(1 + n + i, j)
        Next
        er = er + (correct - predict) ^ 2
    Next
    Cells(out_r, out_c + 3) = "error"
    Cells(out_r, out_c + 4) = er

    se = Cells(out_r + 12, out_c + 2) ' 残差平方和
    aic = n * (Log(se / n)) + 2 * (x2 - x1 + 2)
    Cells(out_r + 1, out_c + 3) = "AIC"
    Cells(out_r + 1, out_c + 4) = aic
End Sub
```

のように作れるかもしれない.

引数の意味は：

・データのうち最初の n 個を教師データとして回帰モデルの予測に使用し，残りの m 個をテストデータとしてクロスバリデーションに使う

・y, x1, x2, out_r, out_c は lm の引数と同じ

で，例えば，

```
Sub hoge()
    predict_check 60, 15, 2, 3, 18, 1, 20
End Sub
```

のようにサブルーチンを作って実行することができる.

また，全ての説明変数に対いて，使うか使わないか，全てのパターンを試して，それぞれの場合の AIC の値や，補正 R2 の値，簡易クロスバリデーションでの誤差の値を列挙するプログラムは

```
Sub brute_force(n, m, y, x1, x2)

    col_max = WorksheetFunction.Max(y, x2) + 3
    row_max = n + m + 3

    Cells(row_max, 1) = "説明変数"
    Cells(row_max, 2) = "AIC"
    Cells(row_max, 3) = "補正 R2"
    Cells(row_max, 4) = "error"

    p = x2 - x1 + 1
    For msk = 1 To 2 ^ p - 1
        sy = col_max
        sx1 = sy + 1
        sx2 = sy
        varname = ""

        Range(Cells(1, sy), Cells(1 + n + m, sy)).Value = Range(Cells(1, y), Cells(1 + n + m, y)).Value
        For i = 0 To p - 1
            If (msk And (2 ^ i)) <> 0 Then
                sx2 = sx2 + 1
                Range(Cells(1, sx2), Cells(1 + n + m, sx2)).Value = Range(Cells(1, x1 +
```

```
i), Cells(1 + n + m, x1 + i)).Value

                    If varname <> "" Then
                        varname = varname + "/"
                    End If
                    varname = varname + Cells(1, x1 + i)
                End If
            Next

            out_r = 1
            out_c = sx2 + 3
            predict_check n, m, sy, sx1, sx2, out_r, out_c

            Cells(row_max + msk, 1) = varname
            Cells(row_max + msk, 2) = Cells(out_r + 1, out_c + 4)
            Cells(row_max + msk, 3) = Cells(out_r + 5, out_c + 1)
            Cells(row_max + msk, 4) = Cells(out_r, out_c + 4)
        Next
End Sub
```

となる.

引数の意味は，lm や predict_check と同じであり，結果はデータの下に追加される.

説明変数の数が増えると計算時間が増えるため，使用する説明変数は少し減らして実行するには,

```
Sub hoge()
    brute_force 60, 15, 2, 3, 8
End Sub
```

とサブルーチンを作成して実行すれば良い.

出力結果をソートすることで，それぞれの基準で一番良いと思われるモデルがわかる.