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=4 7662

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=4 7759

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

	А	В	С	D	E	F	G	Н	I.	J	
1		ビール	東京								
2	2011年1月	475	5.1	回帰分	析					?	×
3	2011年2月	625	7	入力元	Ē					OK	1
4	2011年3月	800	8.1	入力	Y 範囲(<u>Y</u>):	\$	B\$1:\$B\$77	Ť			
5	2011年4月	960	14.5	3.7	X 範囲(X):	s	C\$1:\$C\$77	Ť		キャンセル	
6	2011年5月	730	18.5							∧ルプ(<u>H</u>)	
7	2011年6月	980	22.8		i∕∿l(<u>L)</u> ===== utr(o)	□ 定数	なに 0 を使用(<u>Z</u>	.)			
8	2011年7月	1295	27.3		月忘水準(<u>0</u>)	95	%				
9	2011年8月	1135	27.5	出力才	プション						
10	2011年9月	830	25.1	0-	-覧の出力先(9	<u>5</u>):		Ť			
11	2011年10月	805	19.5	⊛ ≋	f規ワーク シート	·(<u>P</u>):					
12	2011年11月	840	14.9	() 第	f規ブック(<u>W</u>)						
13	2011年12月	1375	7.5	残差							
14	2012年1月	480	4.8		賎差(<u>R</u>) ■淮/レ⇒わち研∺	羊/┬\	□ 残差グラフの	X作成(<u>D</u>) つかた成(I)			
15	2012年2月	610	5.4		長/年16011/27天/	左(土)	□ 電紀沢川担クフ.	7071FD&(<u>1</u>)			
16	2012年3月	810	8.8	正規	確率 ニ担確率ガニコイ						
17	2012年4月	886	14.5		_ スラ玩い田 = キング ブブリ	ντ Ηλα(<u>ΙΝ</u>)					
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.

	А	В	С	D	E	F	G	Н	1
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	रै							
11		自由度	変動	分散	∥された分離	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-值	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
- FB	 の場合で	結果を少	し見ていく	、 」 、 」					

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.

А	В	С	D	E	F	G	Н	1
概要								
回帰	統計							
重相関 R	0.431415							
重決定 R2	0.186119							
補正 R2	0.163821							
標準誤差	213.7403							
観測数	76							
分散分析表	ŧ							
	自由度	変動	分散	∥された分離	有意 F			
回帰	2	762649.4	381324.7	8.346836	0.000544			
残差	73	3335000	45684.94					
合計	75	4097650						
	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	653.1112	77.32065	8.446788	2.04E-12	499.0113	807.211	499.0113	807.211
東京	15.34582	32.42195	0.473316	0.6374	-49.271	79.96268	-49.271	79.96268
京都	-2.15779	29.57313	-0.07296	0.942034	-61.097	56.78137	-61.097	56.78137
	A 概要 回帰 重建定R2 補標準測数 分散分析表 回帰 3 分散分析表 回帰 切片 京都	A B 概要	ABC概要-個冊-回帰-重相関 R0.431415重決定 R20.186119建決定 R20.163821標準誤差213.7403標準誤差213.7403観測数76少市公-個用-自由度変動回帰1618資差公3335000合計-「係数標準誤差切片653.1112東京15.34582哀北157929.57313	ABCD概要ImmediateImmediateImmediate面同学ImmediateImmediateImmediate重相関 R0.431415ImmediateImmediate重決定 R20.186119ImmediateImmediate補正 R20.163821ImmediateImmediate標準誤差213.7403ImmediateImmediate観測数ImmediateImmediateImmediate微微分析ImmediateImmediateImmediate分散分析ImmediateImmediateImmediate分散分析ImmediateImmediateImmediateGImmediateImmediateImmediateGImmediateImmediateImmediateMImmediate<	ABCDE概要 <td< th=""><th>ABCDEF概要II<!--</th--><th>ABCDEFG概要IIIIIII個冊<IIIIIII1日間0.431415IIIIIII重相関0.431415IIIIIII重排定名0.186119IIIIIII補正名0.163821IIIIIII標準誤差213.7403IIIIIII積測数IIIIIIIII分散分析IIIIIIIII月自由度変動分散IIIIIII月生IIIIIIIIII月生1III<td< th=""><th>ABCDEFGH概要International StratementInternational StratementInternational StratementInternational Stratement概要International StratementInternational StratementInternational StratementInternational Stratement面陽0.431415International StratementInternational StratementInternational StratementInternational Stratement面目相関0.431415International StratementInternational StratementInternational StratementInternational Stratement面目相関0.163821International StratementInternational StratementInternational StratementInternational Stratement標準誤登0.163821International StratementInternational StratementInternational StratementInternational Stratement個別数0.163821International StratementInternational StratementInternational StratementInternational Stratement個別数1016International StratementInternational StratementInternational StratementInternational Stratement分散分析表International StratementInternational StratementInternational StratementInternational Stratement内陽International StratementInternational StratementInternational StratementInternational Stratement内陽International StratementInternational StratementInternational StratementInternational Stratement内陽International StratementInternational StratementInternational StratementInternational Stratement内陽</th></td<></th></th></td<>	ABCDEF概要II </th <th>ABCDEFG概要IIIIIII個冊<IIIIIII1日間0.431415IIIIIII重相関0.431415IIIIIII重排定名0.186119IIIIIII補正名0.163821IIIIIII標準誤差213.7403IIIIIII積測数IIIIIIIII分散分析IIIIIIIII月自由度変動分散IIIIIII月生IIIIIIIIII月生1III<td< th=""><th>ABCDEFGH概要International StratementInternational StratementInternational StratementInternational Stratement概要International StratementInternational StratementInternational StratementInternational Stratement面陽0.431415International StratementInternational StratementInternational StratementInternational Stratement面目相関0.431415International StratementInternational StratementInternational StratementInternational Stratement面目相関0.163821International StratementInternational StratementInternational StratementInternational Stratement標準誤登0.163821International StratementInternational StratementInternational StratementInternational Stratement個別数0.163821International StratementInternational StratementInternational StratementInternational Stratement個別数1016International StratementInternational StratementInternational StratementInternational Stratement分散分析表International StratementInternational StratementInternational StratementInternational Stratement内陽International StratementInternational StratementInternational StratementInternational Stratement内陽International StratementInternational StratementInternational StratementInternational Stratement内陽International StratementInternational StratementInternational StratementInternational Stratement内陽</th></td<></th>	ABCDEFG概要IIIIIII個冊<IIIIIII1日間0.431415IIIIIII重相関0.431415IIIIIII重排定名0.186119IIIIIII補正名0.163821IIIIIII標準誤差213.7403IIIIIII積測数IIIIIIIII分散分析IIIIIIIII月自由度変動分散IIIIIII月生IIIIIIIIII月生1III <td< th=""><th>ABCDEFGH概要International StratementInternational StratementInternational StratementInternational Stratement概要International StratementInternational StratementInternational StratementInternational Stratement面陽0.431415International StratementInternational StratementInternational StratementInternational Stratement面目相関0.431415International StratementInternational StratementInternational StratementInternational Stratement面目相関0.163821International StratementInternational StratementInternational StratementInternational Stratement標準誤登0.163821International StratementInternational StratementInternational StratementInternational Stratement個別数0.163821International StratementInternational StratementInternational StratementInternational Stratement個別数1016International StratementInternational StratementInternational StratementInternational Stratement分散分析表International StratementInternational StratementInternational StratementInternational Stratement内陽International StratementInternational StratementInternational StratementInternational Stratement内陽International StratementInternational StratementInternational StratementInternational Stratement内陽International StratementInternational StratementInternational StratementInternational Stratement内陽</th></td<>	ABCDEFGH概要International StratementInternational StratementInternational StratementInternational Stratement概要International StratementInternational StratementInternational StratementInternational Stratement面陽0.431415International StratementInternational StratementInternational StratementInternational Stratement面目相関0.431415International StratementInternational StratementInternational StratementInternational Stratement面目相関0.163821International StratementInternational StratementInternational StratementInternational Stratement標準誤登0.163821International StratementInternational StratementInternational StratementInternational Stratement個別数0.163821International StratementInternational StratementInternational StratementInternational Stratement個別数1016International StratementInternational StratementInternational StratementInternational Stratement分散分析表International StratementInternational StratementInternational StratementInternational Stratement内陽International StratementInternational StratementInternational StratementInternational Stratement内陽International StratementInternational StratementInternational StratementInternational Stratement内陽International StratementInternational StratementInternational StratementInternational Stratement内陽

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.

E	ີ່ 5 ° ∂ ° ∓									グラ	フツール	jma
יידר	イル ホーム 🚦	挿入 ページ	ን レイアウト	数式 デー	-タ 校閲	表示	開発	∿เプ วเเ	P JUST PDF	4 デザイン	, 書式	♀実行し
して ビボ・ テーン	ト おすすめ テ バル ビボットテーブル テーブル	ブル 画像	オンライン 画像			t Xayh t	出 アドインを入 9 個人用アド	、手 イン ・ Bing マップ アドイン	People Graph	ø ₩ • • •		・ ビボットグラ ・
グ	771 ·	×	f _x									
	А	В	С	D	E	F	G	Н	1	J	К	L
1		ビール	東京	京都								
2	2011年1月	475	5.1	2.8		1		 	7		0	+
3	2011年2月	625	7	6.3	_			- 「「「」「」「」	5			
4	2011年3月	800	8.1	6.8	35							
5	2011年4月	960	14.5	12.5	30							
6	2011年5月	730	18.5	19	25							
7	2011年6月	980	22.8	24.1	20							
8	2011年7月	1295	27.3	27.9	020						Ĭ	
9	2011年8月	1135	27.5	28.7	15		•					
10	2011年9月	830	25.1	24.7	10							
11	2011年10月	805	19.5	18.4	5		P .					
12	2011年11月	840	14.9	13.8	0							
13	2011年12月	1375	7.5	6.5	0	5	10	15	20 25	30	35	
14	2012年1月	480	4.8	41	Ŭ			Ŭ				

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.

	А	В	С	D	Е		F	G	н	1	J	K	L
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					E'-	ール			
5	2011年4月	960	14.5	12.5	:	1600							
6	2011年5月	730	18.5	19	:	1400							
7	2011年6月	980	22.8	24.1	:	1200						:	
8	2011年7月	1295	27.3	27.9	:	1000		•		• •••	• •	•	
9	2011年8月	1135	27.5	28.7		800							
10	2011年9月	830	25.1	24.7		600		1.5					
11	2011年10月	805	19.5	18.4		400	-	•					
12	2011年11月	840	14.9	13.8		200							
13	2011年12月	1375	7.5	6.5		0		F 10	15	20	25	20	25
14	2012年1月	480	4.8	4.1			0 :	5 10	15	20	25	30	35
15	2012年2月	610	5.4	4.1	;	系列の	福生			7 X			
16	2012年3月	810	8.8	8.3			(1)						
17	2012年4月	886	14.5	14.2	3	糸列石	(<u>N</u>):						
18	2012年5月	847	19.6	18.8		=jma 玄제 y	(1)\$D\$1 (小店(Y)·		<u> </u>	-)//			
19	2012年6月	998	21.4	22.7		=ima		:77	↑ = 5.1	. 7. 8.1. 1			
20	2012年7月	1193	26.4	27.5	3	, 系列 Y	の値(Y):			., . , ,			
21	2012年8月	1177	29.1	29		=jma	!\$B\$2:\$B\$:77	1 = 47	5, 625, 800,.			
22	2012年9月	810	26.2	25.6						be be be i			
23	2012年10月	818	19.4	18.2				L	OK	キャンセル			
					_								

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.

С A В D E F G н 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 ||された分|| 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 係数 標準誤差 P-値 下限 95% 上限 95% 下限 95.0% 上限 95.0% 16 t 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

これで回帰分析を行うと以下のようになるはずである.

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.

	А	В	С	D	E	F	G	н	1	J	K	
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	
2	2011年1月	475	5.1	回帰分析	ŕ					? ×		0
3	2011年2月	625	7	入力元						OK		0
4	2011年3月	800	8.1	<u>አ</u> ታ	′ 範囲(<u>Y</u>):	\$8	B\$1:\$B\$76	Î				0
5	2011年4月	960	14.5	ג ג ג ג ג	(範囲(X):	\$0	C\$1:\$R\$76	Ť		キャンセル		0
6	2011年5月	730	18.5							∧ルプ(<u>H</u>)		1
7	2011年6月	980	22.8			□ 定数	に 0 を使用(<u>Z</u>)					0
8	2011年7月	1295	27.3		忌水準(<u>○</u>)	95	%					0
9	2011年8月	1135	27.5	出力オフ	ション							0
10	2011年9月	830	25.1	0-5	覧の出力先(<u>S</u>):			Ť				0
11	2011年10月	805	19.5	● 新規	見ワークシート(<u>P</u>)):						0
12	2011年11月	840	14.9	○新規	見ブック(<u>W</u>)							0
13	2011年12月	1375	7.5	残差								0
14	2012年1月	480	4.8		豊(<u>R</u>) 第1ルキわち研業([□残差グラフのℓ	F成(<u>D</u>)				0
15	2012年2月	610	5.4		≢16C4 UC7\$7£(1) L	電光沢川世クラブの	04FDX(<u>1</u>)				0
16	2012年3月	810	8.8		率 見確変ガミコのが	50(N)						0
17	2012年4月	886	14.5			····(<u>14</u>)						0
18	2012年5月	847	19.6									1

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.

	А	В	С	D	E	F	G	Н	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	ŧ							
11		自由度	変動	分散	された分	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○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 \rightarrow オプション \rightarrow アドイン \rightarrow 設定 \rightarrow 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.

	A	В	С	D	E	F	G	Н	Ι	J	K	L	M	N	0	P	(ב	R	S	;	Т		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		1			225625		60573216	
3	2011年2月	625	7	6.3	8.1	2			0	. (0 (0 (0	0	0		1			390625			
4	2011年3月	800	8.1	6.8	14.5	3	ענע	(-の/(:	58-9-													;	×	640000			
5	2011年4月	960	14.5	12.5	18.5	4																		921600			
6	2011年5月	730	18.5	19	22.8	5		目的セ	レの設定	:(<u>T</u>)				\$W\$3	2							1		532900			
7	2011年6月	980	22.8	24.1	27.3	e		目標値		大値(<u>M)</u>) 最小(<u>∎(N</u>	○ 指定	値:(⊻)		0							960400			
8	2011年7月	1295	27.3	27.9	27.5	7			の赤声	·(P)														1677025			
9	2011年8月	1135	27.5	28.7	25.1	8	ſ	\$C\$78	:\$5\$78	·(묘)												Ť		1288225			
10	2011年9月	830	25.1	24.7	19.5	9																		688900			
11	2011年10月	805	19.5	18.4	14.9	10	ſ	制約条	件の対象	≹:(<u>U</u>)														648025			
12	2011年11月	840	14.9	13.8	7.5	11														追	助(A)			705600			
13	2011年12月	1375	7.5	6.5	4.8	12														索	Ē更(<u>C</u>)			1890625			
14	2012年1月	480	4.8	4.1	5.4	13														240	(時令(D))			230400			
15	2012年2月	610	5.4	4.1	8.8	14														н	168 (<u>D</u>)			372100			
16	2012年3月	810	8.8	8.3	14.5	15														ৰুশ্ব	リセット(<u>R)</u>		656100			
17	2012年4月	886	14.5	14.2	19.6	16														+	- 10+	(1.)		784996			
18	2012年5月	847	19.6	18.8	21.4	17					1- 4 Whi		•					~	1	売み込	め/保付	•(<u>L</u>)		717409			
19	2012年6月	998	21.4	22.7	26.4	18		刑∦	JU)/2012	EBX 23	中夏欧	-9 S(K)						_					996004			
20	2012年7月	1193	26.4	27.5	29.1	19		解決方: (E)	法の選択	रः (GRG ₿	線形							~	7	プション	(<u>P</u>)		1423249			
21	2012年8月	1177	29.1	29	26.2	20																		1385329			
22	2012年9月	810	26.2	25.6	19.4	21		解決が	フ)法 ポンコド2月1日	∞ 5 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7	Mute	_ 8885(-			₽% T NS	5、9	9∓%⊅=	= ਰਾਮ	0. K.	8885	-/+ I D	257		656100			
23	2012年10月	818	19.4	18.2	12.7	22		レックフ	エンジン	へ滑ら	かではな	い非線	形を示	すソルバ	一問題	にはエブ	ボリュー	・ショナ	FU- 1	ロシジン	を選択	575		669124			
24	2012年11月	865	12.7	11.1	7.3	23		たさい。																748225			
25	2012年12月	1317	7.3	5.4	5.5	24				_									_					1734489			
26	2013年1月	485	5.5	3.9	6.2	25		NK	ブ(土)							1	解決(<u>S</u>)			閉じる	(<u>0</u>)		235225			
27	2013年2月	610	6.2	4.5	12.1	26			0	()	0 0)	0 0) ()	0	0	0	_	1		-	372100			
28	2013年3月	790	12 1	97	15.2	27		n (1 1	()	n r	1	0 (n r)	0	0	0		1			624100			

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	В	С	D	E	F	G	н	I	J	K	L	M	N	0	P	Q	R	S	Т	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
З	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 \wedge 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, sy)).
m, y)).Value
                                         For i = 0 To p - 1
                                                              If (msk And (2 \land i)) \iff 0 Then
                                                                                   sx2 = sx2 + 1
                                                                                   Range(Cells(1, sx2), Cells(1 + n + m, sx2)).Value = Range(Cells(1, x1 + m, sx2)).Va
```

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.