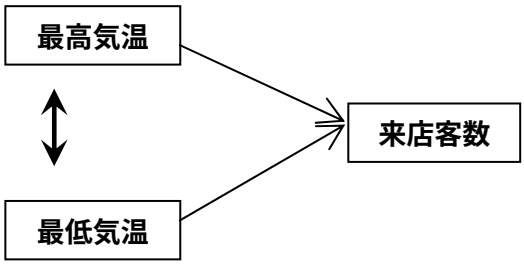
12 重回帰分析

単回帰分析と同様の考え方で，目的変数を予測する説明変数を複数想定する場合が重回帰分析である．

サンプルデータは，あるアイスクリームショップの 20 日間の来店客数と，各日の最高気温と最低気温である．来店客数に，最高気温と最低気温は影響力をもっているだろうか？ これを重回帰分析で検討してみることにしよう．

12.1 重回帰分析の考え方

【来店客数に関する重回帰モデル】



ここで単純に目的変数と説明変数間の相関係数を求めると，

最高気温－来店客数：0.771

最低気温－来店客数：0.138

となる．これだけを見ると，最高気温が高いときに来店客数が増える，という関係は強そうだが，最低気温は来店客数とは関係がなさそう (＝予測力がなさそう) に見えるが，どうだろうか．

日数	最高気温	最低気温	来店客数
1	33	21	382
2	30	20	324
3	34	22	338
4	34	22	317
5	35	25	341
6	32	23	360
7	34	23	339
8	32	21	329
9	28	22	218
10	35	22	402
11	33	23	342
12	28	20	205
13	32	22	368
14	28	23	196
15	28	19	304
16	30	23	294
17	29	21	275
18	32	24	336
19	31	20	384
20	35	24	368

しかし，この 3 変数間にはもう一つの相関関係が存在する．

最高気温－最低気温：0.537

である．つまり，客数だけではなく，最低気温も最高気温の影響を受けていると考えることができるだろう．

最低気温と来店客数との間の純粋な関係を導き出すためには，最高気温から両者への影響力を取り除いた上で，両者の相関係数を求める必要がある．このような相関係数のことを偏相関係数とよぶ．

【偏相関の求め方】

SAS で偏相関を求めるためには，PROC CORR プロシージャで，partial ステートメントを用いて取り除く変数を指定すればよい (上のデータは講義用サイトから参照可能) .

```
proc corr;
  var min guest;
  partial max;
run;
```

偏相関係数 (Pearson), N = 20
帰無仮説 Rho=0 **に対する** Prob > |r|

	min	guest
min	1.00000	-0.51511 0.0240
guest	-0.51511 0.0240	1.00000

最高気温の影響を取り除いた上での最低気温と来店客数の偏相関係数は -0.515 で，有意に 0 とは異なる．

このことから，最高気温が同じであれば，最低気温が低い方が来店客数は多い，という関係があるらしい，ということが予測できる．ちなみに，最低気温の影響を取り除いた上での最高気温と来店客数の偏相関係数は 0.835 となる．最低気温が一定なら，最高気温が高い方が来客は多いようだ．

重回帰分析の場合，単回帰分析とは異なり，単なる「目的変数－説明変数」間の相関係数ではなく，この偏相関係数にもとづいてそれぞれの説明変数と目的変数（ここでは「最高気温 x と来店客数 y」と「最低気温 x と来店客数 y」）の関係を示す偏回帰直線の傾きを求めることになる．この傾きのことを，偏回帰係数とよぶ．つまり偏回帰係数というのは，当該説明変数が，他の説明変数をすべて「一定」にした場合に，目的変数に対してもつ影響力の大きさを示す指標であるといえる．

なお，前回述べた回帰係数の場合と同様に，この偏回帰係数は，それぞれの説明変数がとる値の範囲（データの散らばり）や単位に依存した値をとる．そこで，回帰分析では，全ての変数を平均 0，分散 1 になるように標準化した標準偏回帰係数を算出することが必要となる．この標準偏回帰係数は．当該説明変数が，他の説明変数をすべて「一定」にした場合に，目的変数に対してもつ相対的な影響力の強さを示す指標であるといえる．重回帰分析の結果を記述する場合は，各説明変数が絶対的に有意なものであるかどうかを検討することと同様に，どの説明変数の影響が大きいかを相対的に見比べることにも関心がある場合が多いので，この標準偏回帰係数が有力な指標となる．

12-2 SAS による重回帰分析

前回の単回帰分析と同じく，PROC REG プロシージャを用いて重回帰分析をおこなう．モデルステートメントの右辺に，複数の説明変数を列記すればよい．

```
data icecream;
input day max min guests;
cards;
1   33   21   382
2   30   20   324
3   34   22   338
4   34   22   317
5   35   25   341
6   32   23   360
7   34   23   339
8   32   21   329
9   28   22   218
10  35   22   402
11  33   23   342
12  28   20   205
13  32   22   368
14  28   23   196
15  28   19   304
16  30   23   294
17  29   21   275
18  32   24   336
19  31   20   384
20  35   24   368
;
proc reg;
  model guests=max min / stb;
  print r ;
run;
quit;
```

Dependent Variable: guests

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	45509	22755	20.07	<.0001
Error	17	19272	1133.67618		
Corrected Total	19	64782			

Root MSE	33.67011	R-Square	0.7025
Dependent Mean	321.10000	Adj R-Sq	0.6675
Coeff Var	10.48586		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	-77.18642	118.56103	-0.65	0.5237	0
max	1	22.72126	3.63566	6.25	<.0001	0.98014
min	1	-14.58371	5.88556	-2.48	0.0240	-0.38862

アイスクリームショップへの来店客数には，当日の最高気温，最低気温ともに有意な説明力を持っており，（最低気温が一定であるとすれば）最高気温が高い方が，あるいは，（最高気温が一定であるとすれば）最低気温が低い方が，多くの来客があると予測することができる．

また，来店客数を予測する回帰式は，

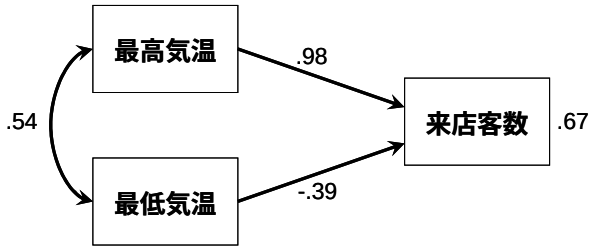
来店客数 guests＝22.721×最高気温 max－14.584×最低気温 min－77.186

となる．

12-3 重回帰モデルの解釈

出力結果から，以下の順で結果の解釈をおこなう．

1) **モデル全体の有意性検定** (Analysis of Variance **欄の** *F* **検定結果**)
   今回検討したモデルが null model (**どの変数も効果をもたないというモデル**) **と有意に異なる確率**
2) **予測力の確認** (**決定係数は意味のある大きさか**； Adj R-Sq)
   じゅうぶんに大きな値であるか，自由度を調整していない決定係数(R-Square)**との差は小さいかを検討する**
3) **偏回帰係数は影響力があるか** (**各説明変数の偏回帰係数の** *t* **検定結果**)
   今回投入した説明変数の偏回帰係数がゼロと有意に異なる確率
4) **どの独立変数の説明力が大きいか**
   独立変数の測定単位，分散が同じであれば，偏回帰係数を基準に「大きさ」で判断してよい．いずれかが異なる場合は，標準偏回帰係数を基準に「強さ」を判断する．ただし，独立変数の影響を比較することはかなり難しいので，有意なものと有意でないものが識別できる以外は，とりあえず「見定める」程度でよい．



12-4 多重共線性(multicolinearlity；通称マルチコ)

重回帰分析は，独立変数同士に強い相関がある場合に適切に分析が実行できないことがある．このことを「多重共線性」という．多重共線性が生じると， SAS **の出力結果は不自然なものになる** (**標準偏回帰係数が** 1 **を超える，決定係数が非常識的に** 1 **に近い値をとる，不当に大きな，あるいは解釈できないような推定結果が得られる等々**)**．多重共線性が生じるのを防ぐために，独立変数として投入する変数間の相関係数をあらかじめ求めておき，非常に大きな(めやすとしては|*r*|=0.9 以上)ものがないかどうか確かめておこう．もしそのような変数群があれば，重要度が低い変数を除去した上で重回帰分析を実施する必要がある．**

また，こうした多重共線性のチェック方法としては，モデルステートメントに VIF (Variance Inflation Factor; 分散拡大係数) オプションをつけて実行し，VIF **の値を検討するものがある．常識的には，** VIF **の値は** 5 **以下程度であるが，このだいたいの基準を大きく逸脱** (**例えば** VIF>10) **しているような変数は，マルチコを引き起こしている可能性が高いので，重要な変数でないのなら，独立変数から除外した方がよい．**

```
proc reg;
  model guests=max min / stb vif;
  print r ;
run;
quit;
```

ただしこうした判断は，絶対的な基準があるわけではなく，ケースやデータによってさまざまに異なるので，ベストのモデルが何であるかを判断するためには，それぞれの変数や相互の関連のもつ特徴なども加味した柔軟な対応が必要である．

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-77.18642	118.56103	-0.65	0.5237	0
max	1	22.72126	3.63566	6.25	<.0001	1.40555
min	1	-14.58371	5.88556	-2.48	0.0240	1.40555

12-4 適切な変数選択

重回帰分析において，説明変数として投入するいくつかの変数群から，目的変数に有意な効果をもつ変数のみを選び出して投入する方法がある．適切な変数選択をおこなうことは，結果記述のシャープさを増すためにも重要であるので，特に想定される説明変数群が多い (精選しにくい) 状況の場合は変数選択オプションをつけて実行することを勧める．

変数選択の方法にはいくつかあるが，例えばステップワイズ法 SELECTION=STEPWISE がある．[1]この方法は「逐次変数選択法」とも呼ばれ，目的変数に大きく影響する説明変数をソフトウェアが自動的に選択して回帰式を求めてくれる．この SELECTION オプションも，VIF オプションと同様に，モデルステートメントに付け加えるとよい．

```
proc reg;
  model guests=max min / stb vif selection=stepwise;
  print r ;
run;
quit;
```

12-5 残差分析

重回帰分析において (回帰式によって求められる) 予測値と実測値 (実際のデータ) の乖離について検討する残差分析をおこなうことが求められる場合がある．この残差分析をおこなうためのオプションが print ステートメントにおける r オプションである．このオプションに基づく出力の結果，理論的に求められた回帰直線から離れたプロット点が存在する場合には，重回帰モデルの妥当性が疑わしいと判断される．例えば標準化残差 (Student Residual)が 2 以上の場合に，外れ値の可能性を疑うことが多い．いくつか大きく逸脱したケースがある場合には，測定ミス，データ入力ミスなどの可能性もあるので，そのような回帰分析によって得られた結果は，より注意深く検討すべきであるし，必要があれば当該データを特定し，削除するなどの対応をおこなう方がよい場合もある．

課題

単回帰分析の例題とした出版ページ数と原稿ページ数の関係に，いくつかの説明変数を加えたデータについて，重回帰分析をおこなえ．特に以下のことに注意しながら，出版ページ数を説明するもっとも適切なモデルを探索し，結果を報告せよ．プログラムのサンプルは，講義用ウェブサイトから参照できる．

決定係数のチェック

Adj R-Sq の値と，R-Square の差を検討せよ

多重共線性のチェック

独立変数間の相関はどの程度か

VIF の値は大きくないか

変数選択 (ステップワイズ法 stepwise と総当たり法 rsquare を用いて出力を見比べ，読み取ってみよう)

統計ソフトはどの変数を組み込んだモデルを最適と判断するか

---

[1] その他の変数選択法として，総当たり法(rsquare；p 個の独立変数のすべての組み合わせ(2p-1)個を検討)，変数増加法(forward；影響力の大きい変数から順に追加)，変数減少法(backward；全変数モデルから影響力の小さい変数を除去)がある．使い分けに明確なルールはないが，従属変数に影響を及ぼす変数を見定めたい (探索的な検討) の場合，さまざまな選択法を試してみるとよいだろう．なお，変数選択法を用いた場合は，結果記述の際にどの方法を使ったか明記すること．