データ科学理論特論

客員教授　藤越　康祝

データ解析のための統計理論を、典型的な実データ、および、種々の統計モデルを通して講義する。モデルとしては、直線回帰モデル、一元配置モデル、回帰モデル、一般化線形モデル、自己回帰モデル、成長曲線モデル、などを取り上げる。統計的概念としてく最小二乗法、最尤法とその基本的性質、有意性検定法、同時推測法、予測法、などの基礎を説明する。また、各種モデルにおける変数の選択や次元の推定をAIC基準やCp基準などのモデル選択を利用して行う方法についても講義する。さらに、複雑な推測法を数値的に評価するための方法であるシミュレーション法・ブートストラップ法や、データのグラフィカル表現についても講義する。

例えば、モデル選択の例として、セメントの硬化に関する例を取り上げる。目的変数 y はセメント 1g 当り放出する熱量であり、説明変数は４種数のカルシウムの量 x_1, x_2, x_3, x_4 である。このとき、y を予測するのに４つの説明変数すべてを用いるのは賢明でないかもしれない。変数をすべて取り入れて重回帰式

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

を作れば残差平方和は小さくなり、確かに重相関係数は大きくなる（すなわち、予測の精度がよいように見える）。しかしながら、この予測式の安定性は悪くなる。４つの説明変数すべてを用いた場合と、そのうちの一部を用いた場合とを比較して、予測の精度がそれほど変わっていなければ、説明変数の一部を用いたときの重回帰式で十分間に合い、またその方が重回帰式は安定している。このような変数を如何に選べばよいかが、重回帰式の変数選択問題である。

セメントの硬化に関するデータ

j	x_{j1}	x_{j2}	x_{j3}	x_{j4}	y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	29	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

変数選択法としては、変数増加法、変数減少法、変数増減法、変数減増法等が知られている。また、残差平方和とモデルに含まれる変数の数の両方を小さくする基準 AIC, Cp などがある。変数減少法は４つの変数からなる完全モデルから出発し、１つの変数を除いたときの部分的 F 値を求め、これが最小になる変数をまず除去する。除去した結果、３つの変数をもつモデルが得られる。このモデルにおいて、１つの変数を除いた部分的 F 値を考え、次にこれらが最小になる変数を除去する。この操作で、最小になる F 値が有意であれば、その変数は除去しないで、そこで操作を打ち切る。

完全モデル

変数	b_i	$s(b_i)$	$F(\beta_i = 0)$
x_1	1.551	0.745	4.33
x_2	0.510	0.724	0.50
x_3	0.102	0.755	0.02
x_4	-0.144	0.710	0.04
切片	62.418		

選ばれたモデル

変数	b_i	$s(b_i)$	$F(\beta_i = 0)$
x_1	1.468	0.121	146.52
x_2	0.662	0.064	208.53
切片	52.577		

b_i : β_iの推定値

$s(b_i)$: b_iの標準誤差

４変数のさいに示された x_3, x_4 のＦ値は極めて小さい。実際には、x_3 が除去され、次に x_4 が除去されて、最終結果として下記の予測式を得る。

$$\hat{y} = 52.577 + 1.468 x_1 + 0.662 x_2$$

一方、すべての変数の組を考え、AIC, Cp が最小になるモデルを求めると、同じ結果になる。

(文責　数学専攻　修士２年　小椋　透)