第 13 章

一般化線型モデル入門

13.1 一般化線型モデルとは？

第 11 章で説明したように，普通の量的変数の間の線型回帰を一般化すれば，t 検定，分散分析，共分散分析，回帰分析，重回帰分析，ロジスティック回帰分析，正準相関分析など多くの分析方法を共通の数学モデルで扱うことができる。このモデルは一般化線型モデルと呼ばれる。英語では Generalized Linear Model といい，R での関数名も `glm()` である[*1]。

一般化線型モデルは，基本的には，$Y = \beta_0 + \beta X + \varepsilon$ という形で表される（Y が従属変数群，X が独立変数群（及びそれらの交互作用項），β_0 が切片群，β が係数群，ε が誤差項である）。係数は最小二乗法または最尤法で数値的に求める。以下，先にあげたいくつかの分析が，どのように一般化線型モデルを特殊化したものなのかを説明し，その中で重回帰分析と共分散分析について若干の補足説明を加える。

13.2 変数の種類と数の違いによる線型モデルの分類

以下のように整理すると，t 検定，分散分析，回帰分析といった分析法が，すべて一般化線型モデルの枠組みで扱えることがわかる。

[*1] 線型は linear の訳で，一次結合という意味なのだが，漢字としては線形と書かれることもある。厳密な区分はないように思われるが，`glm()` の場合は「型」の字を使う方が普通のようである。なお，一般化線型モデルのうち，ある条件を満たすものを一般線型モデル (General Linear Models) と呼び，SAS の PROC GLM はこれに当たる。

分析名	従属変数 (Y)	独立変数 (X)
t 検定（注 1）	量的変数 1 つ	2 分変数 1 つ
一元配置分散分析	量的変数 1 つ	カテゴリ変数 1 つ
多元配置分散分析	量的変数 1 つ	カテゴリ変数複数
回帰分析	量的変数 1 つ	量的変数 1 つ
重回帰分析	量的変数 1 つ	量的変数複数（注 2）
共分散分析	量的変数 1 つ	（注 3）
ロジスティック回帰分析	2 分変数 1 つ	2 分変数，カテゴリ変数，量的変数複数
正準相関分析	量的変数複数	量的変数複数

（注 1）Welch の方法でない場合。

（注 2）カテゴリ変数はダミー変数化。

（注 3）2 分変数 1 つと量的変数 1 つの場合が多いが，「2 分変数またはカテゴリ変数 1 つまたは複数」と「量的変数 1 つまたは複数」を両方含めば使える。

たとえば，建物の型の変数 (BD) を集合住宅 1，一戸建て 2 とした場合の，東京のとある大学の学生実習で測定した水道水質の総硬度 (HARD) の平均に，建物の型によって差があるかどうかを検定したいとする。

等分散性を仮定すれば，R では，

```
BD <- c(1,1,1,1,1,1,2,2,1,1,2,1,1,2,1,1,2,2,1,1,1,1,1,2,1,1,2,1,1,2,1,2,1,1)
HARD <- c(88.280, 103.500, 119.600, 96.210, 109.340, 100.500, 81.390, 75.715,
112.880, 101.150, 84.400, 102.900, 65.000, 97.445, 101.850, 79.100, 103.620,
69.270, 97.090, 101.150, 89.820, 108.560, 98.810, 103.620, 85.940, 89.230,
69.300, 101.150, 101.150, 73.070, 62.695, 148.590, 93.080, 103.500)
```

としてデータを定義した後，`t.test(HARD ~ BD, var.equal=T)` とすることによって，以下の結果が得られる。

```
Two Sample t-test
data:  HARD by BD
t = 0.8843, df = 32, p-value = 0.3831
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.444719 18.867802
sample estimates:
mean in group 1 mean in group 2
96.35354        90.64200
```

一般化線型モデルを使って，建物の型を独立変数として総硬度を従属

変数としたモデルの当てはめをしてみるには，R では，データ定義後に，summary(glm(HARD ~ BD)) とすればよい。以下の結果が得られる。

```
Call:
glm(formula = HARD ~ BD)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-33.659   -8.957    3.301    7.061   57.948

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  102.065      8.861  11.518 6.41e-13 ***
BD            -5.712      6.459  -0.884    0.383
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 294.4717)

    Null deviance: 9653.4  on 33  degrees of freedom
Residual deviance: 9423.1  on 32  degrees of freedom
AIC: 293.72

Number of Fisher Scoring iterations: 2
```

Coefficients:の BD のところを見ると，t value が -0.884，その有意確率が 0.383 となっていて，t 検定の結果と一致している（t 値の符号が違うが，t 分布は左右対称なので両側検定では符号が違っても同じ意味）ことがわかる。

この場合は，当然のことながら，普通の線型モデルでも同じ結果が得られるし，分散分析でも同じ結果となる。つまり，t 検定は分散分析の特殊な場合ということができるし，分散分析は線型モデルの特殊な場合ということができるし，線型モデルは一般化線型モデルの特殊な場合（当然だが）ということができる。

13.3 重回帰分析

複数の独立変数を同時にモデルに投入することにより，従属変数に対する，他の影響を調整した個々の変数の影響をみることができる。

重回帰分析は，何よりもモデル全体で評価することが大切である。たとえば，独立変数が年齢と体重と一日当たりエネルギー摂取量，従属変数が血圧というモデルを立

てれば，年齢の偏回帰係数（または偏相関係数または標準化偏回帰係数）は，体重と一日当たりエネルギー摂取量の影響を調整した（取り除いた）後の年齢と血圧の関係を示す値だし，体重の偏回帰係数は年齢と一日当たりエネルギー摂取量の影響を調整した後の体重と血圧の関係を示す値だし，一日当たりエネルギー摂取量の偏回帰係数は，年齢と体重の影響を調整した後の一日当たりエネルギー摂取量と血圧の関係を示す値である。独立変数が年齢と体重で従属変数が血圧である場合の年齢の偏回帰係数は，独立変数に一日当たりエネルギー摂取量も入っている場合の年齢の偏回帰係数とは異なる。

モデル全体としてのデータへの当てはまりは，重相関係数の2乗（決定係数）や，AIC で評価する。

偏回帰係数の有意性検定は，偏相関係数がゼロである確率を t 検定によって求める。1つの重回帰式の中で，相対的にどの独立変数が従属変数（の分散）に対して大きな影響を与えているかは，偏相関係数の二乗の大小によって評価するか，または標準化偏回帰係数によって比較することができる。しかし，原則としては，別の重回帰モデルとの間では比較不可能である。

たくさんの独立変数の候補からステップワイズ法によって比較的少数の独立変数を選択することが良く行われる。しかし，モデル全体で評価するという観点からは，あまり薦められない。数値以外の根拠により投入する変数を決めて，各々の偏回帰係数（または偏相関係数）が有意であるかないかを見る方が筋がよい。十分な理由があれば，有意でない変数も含めた重回帰式を作っても良い。

しかし，数値以外の根拠が薄い場合もあるし，偏回帰係数が有意でない（偏相関係数がゼロであるという帰無仮説が成り立つ確率が5%より大きい）変数を重回帰モデルに含めることを嫌う立場もある。したがって，数値から最適なモデルを求める必要もありうる。そのためには，独立変数が1個の場合，2個の場合，3個の場合，……，のそれぞれについてすべての組み合わせの重回帰モデルを試して，もっとも重相関係数の二乗が大きなモデルを求めて，独立変数が n 個の場合が，$n-1$ 個の場合のすべての変数を含むならば尤度比検定を行って，尤度が有意に大きくならないところまでの $n-1$ 個を独立変数として採用するのが良い。SAS では PROC REG の MAXR というオプションで可能である。

13.4 共分散分析

典型的には，$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_{12}X_1X_2+\varepsilon$ というモデルになる。2分変数 X_1 によって示される2群間で，量的変数 Y の平均に差があるかどうかを比べるのだが，Y が量的変数 X_2 と相関がある場合に（このとき X_2 を共変量と呼ぶ），X_2

と Y の回帰直線の傾き (slope) が X_1 の示す 2 群間で差がないときに，X_2 による影響を調整した Y の修正平均 (adjusted mean; 調整平均ともいう) に，X_1 の 2 群間で差があるかどうかを検定する。

R では，X_1 を示す変数名を C (注：C は factor である必要がある)，X_2 を示す変数名を X とし，Y を示す変数名を Y とすると，summary(glm(Y~C+X)) とすれば，X の影響を調整した上で，C 間で Y の修正平均 (調整平均) が等しいという帰無仮説についての検定結果が得られる (C2 と表示される行の右端に出ているのがその有意確率である)。ただし，この検定をする前に，2 本の回帰直線がともに有意にデータに適合していて，かつ 2 本の回帰直線の間で傾き (slope) が等しいかどうかを検定して，傾きが等しいことを確かめておかないと，修正平均の比較には意味がない。そこで，まずたとえば，summary(lm(Y[C==1]~X[C==1]);summary(lm(Y[C==2]~X[C==2]) として 2 つの回帰直線それぞれの適合を確かめ，summary(glm(Y~C+X+C*X)) として傾きが等しいかどうかを確かめなければならない。傾きが有意に違っていることは，C と X の交互作用項が有意に Y に効いていることと同値なので，Coefficients の C2:X と書かれている行の右端を見れば，「傾きが等しい」を帰無仮説とした場合の有意確率が得られる。そもそも回帰直線の適合が悪ければその独立変数は共変量として考慮する必要がないし，傾きが違っていれば群分け変数と独立変数の交互作用が従属変数に関して有意に影響しているということなので，2 群を層別して別々に解釈する方が良い。

参考までに数式でも説明しておく。いま，C で群分けされる 2 つの母集団における，(X, Y) の間の母回帰直線を，$y = \alpha_1 + \beta_1 x$，$y = \alpha_2 + \beta_2 x$ とすれば，次の 2 つの仮説が考えられる。まず傾きに差があるかどうか？　を考える。つまり，$H_0 : \beta_1 = \beta_2$，$H_1 : \beta_1 \neq \beta_2$ である。次に，もし傾きが等しかったら，y 切片も等しいかどうかを考える。つまり，$\beta_1 = \beta_2$ のもとで，$H_0' : \alpha_1 = \alpha_2$，$H_1' : \alpha_1 \neq \alpha_2$ を検定する。各群について，X と Y の平均と変動と共変動を出しておけば[*2]，仮説 H_1 のもとでの残差平方和

$$d_1 = SS_{Y1} - (SS_{XY1})^2/SS_{X1} + SS_{Y2} - (SS_{XY2})^2/SS_{X2}$$

と仮説 H_0 のもとでの残差平方和

$$d_2 = SS_{Y1} + SS_{Y2} - (SS_{XY1} + SS_{XY2})^2/(SS_{X1} + SS_{X2})$$

を計算して $F = (d_2 - d_1)/(d_1/(N-4))$ が H_0 のもとで第 1 自由度 1，第 2 自由度 $N-4$ の F 分布に従うことを使って傾きが等しいかどうかの検定ができる。H_0 が棄

[*2] 標本サイズ $N1$ の第 1 群に属する x_i, y_i について，$E_{X1} = \sum x_i/N1$，$SS_{X1} = \sum (x_i - E_{X1})^2$，$E_{Y1} = \sum y_i/N1$，$SS_{Y1} = \sum (y_i - E_{Y1})^2$，$E_{XY1} = \sum x_i y_i/N1$，$SS_{XY1} = \sum (x_i y_i - E_{XY1})^2$。第 2 群も同様。

却されたときは，$\beta_1 = SS_{XY1}/SS_{X1}$，$\beta_2 = SS_{XY2}/SS_{X2}$ として別々に傾きを推定し，y 切片 α もそれぞれの式に各群の平均を入れて計算できる。H_0 が採択されたときは，共通の傾き β を，$\beta = (SS_{XY1} + SS_{XY2})/(SS_{X1} + SS_{X2})$ として推定する。この場合はさらに y 切片が等しいという帰無仮説 H_0' のもとで全部のデータを使った残差平方和 $d_3 = SS_Y - (SS_{XY})^2/SS_X$ を計算して，$F = (d_3 - d_2)/(d_2/(N-3))$ が第 1 自由度 1，第 2 自由度 $N-3$ の F 分布に従うことを使って検定できる。H_0' が棄却された場合は各群の平均を共通の傾きに代入すれば各群の切片が求められるし，採択されたら，要するに 2 群間に差がないということになるので，2 群を一緒にして普通の単回帰分析をしていいことになる。

例題

下表は，都道府県別のデータで，1990 年の 100 世帯あたり乗用車台数 (CAR1990) と，1989 年の人口 10 万人当たり交通事故死者数 (TA1989) と，1985 年の国勢調査による人口集中地区居住割合 (DIDP1985) である。REGION の 1 は東日本，2 は西日本を意味する。東日本は西日本よりも，人口集中地区居住割合を調整しても乗用車保有台数が多いといえるか？

PREF	REGION	CAR1990	TA1989	DIDP1985
Hokkaido	1	86	11.6	66.7
Aomori	1	78.9	9.5	42.2
Iwate	1	86.6	9.7	27.5
Miyagi	1	92.7	7.9	50.7
Akita	1	90.3	8.1	31.2
Yamagata	1	104.7	7.1	36.7
Fukushima	1	102.7	12.1	33.6
Ibaraki	1	120.7	16.4	29.2
Tochigi	1	122.2	16.5	35.1
Gunma	1	123.9	11.5	38.2
Saitama	1	88.7	7.3	71.7
Chiba	1	86.4	8.8	65
Tokyo	1	58.2	4.1	97.1
Kanagawa	1	75.5	7.2	89.1
Niigata	1	93.2	11.1	42.6
Toyama	1	113	11.1	37.9

PREF	REGION	CAR1990	TA1989	DIDP1985
Ishikawa	1	99.1	9.5	46.4
Fukui	1	109.4	14.7	35.9
Yamanashi	1	112.8	13.8	31.2
Nagano	1	110.9	9.6	31.1
Gifu	1	119.7	12	36.8
Shizuoka	1	107.5	10.5	51.5
Aichi	1	107.2	8.2	67.2
Mie	1	106.7	13.7	38
Shiga	2	104.4	14.5	29.1
Kyoto	2	75.5	8.9	79.5
Osaka	2	62.8	5.9	93.8
Hyogo	2	75.6	8.9	71.7
Nara	2	86	9.3	52.7
Wakayama	2	83	11.6	42.3
Tottori	2	92.1	11.8	26.2
Shimane	2	86.9	9.9	23.4
Okayama	2	95.7	11.3	33.9
Hiroshima	2	79.6	9.7	58.5
Yamaguchi	2	84.4	11.5	44
Tokushima	2	90.7	10.9	27.4
Kagawa	2	89.8	14.3	32.3
Ehime	2	72.3	10.9	43.1
Kochi	2	74.9	11.3	38.4
Fukuoka	2	82.3	8	63.3
Saga	2	97.4	12.8	27.6
Nagasaki	2	69.3	5.9	41.6
Kumamoto	2	87.3	8.5	36.6
Oita	2	82.5	8.7	40.4
Miyazaki	2	85.7	7.4	39
Kagoshima	2	70.5	7.3	36.3
Okinawa	2	100.3	7.6	56.5

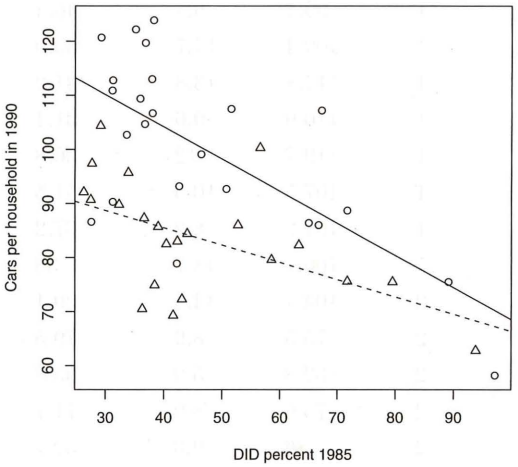

図 13.1 交通事故件数と世帯当たりの自家用車保有台数の関係の東日本と西日本の比較

人口集中地区[*3]人口割合が高い都道府県ほど人がまとまって住んでいるわけだから，先験的に，そういう都道府県ほどマイカー保有率は低くて済みそうだと思われる。したがって，人口集中地区人口割合によってマイカー保有率を調整しなくては，それ以外の要因（たとえば，公共交通機関の整備の割合や，自動車産業の発達の度合い，ディーラーの営業活動の熱心さ，平均世帯規模，郊外型大型店舗の展開の度合い，道路政策，等々）による東日本と西日本のマイカー保有率への影響を評価できないことになる。

東日本を○で，西日本を△でプロットし，東日本の回帰直線を実線，西日本の回帰直線を点線で追加すると，図 13.1 のようになる。この図を描く R のプログラムは，たとえば

```
x <- read.table("anacova.dat")
attach(x)
```

[*3] 1 km^2 当たりの人口密度が 4,000 人以上の集合地区で，かつ合計人口が 5,000 人以上の地区をいう。

としてデータを読み込んでから，

```
plot(CAR1990[REGION==1]~DIDP1985[REGION==1],pch=1,
xlab='DID percent 1985',ylab='Cars per household in 1990')
points(CAR1990[REGION==2]~DIDP1985[REGION==2],pch=2)
abline(lm(CAR1990[REGION==2]~DIDP1985[REGION==2]),lty=2)
abline(lm(CAR1990[REGION==1]~DIDP1985[REGION==1]),lty=1)
```

とすればよい。

R での共分散分析の手順は，まず

```
summary(lm(CAR1990[REGION==1] ~ DIDP1985[REGION==1]))
summary(lm(CAR1990[REGION==2] ~ DIDP1985[REGION==2]))
```

とする。得られる結果

```
Call:
lm(formula = CAR1990[REGION == 1] ~ DIDP1985[REGION == 1])

Residuals:
     Min       1Q   Median       3Q      Max
-24.9808  -5.1307   0.9493   8.2336  19.2190

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           127.9283     6.7699  18.897 4.35e-15 ***
DIDP1985[REGION == 1]  -0.5945     0.1333  -4.459 0.000197 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.24 on 22 degrees of freedom
Multiple R-Squared: 0.4747,Adjusted R-squared: 0.4508
F-statistic: 19.88 on 1 and 22 DF,  p-value: 0.0001967

Call:
lm(formula = CAR1990[REGION == 2] ~ DIDP1985[REGION == 2])

Residuals:
     Min       1Q   Median       3Q      Max
-16.1869  -3.3935   0.2297   3.4338  20.0706
```

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            98.2912     5.0750  19.368 7.12e-15 ***
DIDP1985[REGION == 2]  -0.3197     0.1047  -3.053  0.00604 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.904 on 21 degrees of freedom
Multiple R-Squared: 0.3075,Adjusted R-squared: 0.2745
F-statistic: 9.323 on 1 and 21 DF,  p-value: 0.006037
```

から，これらの回帰式が両方とも有意にデータに適合していることがわかる。次に，

```
summary(glm(CAR1990 ~ as.factor(REGION)+DIDP1985
+as.factor(REGION)*DIDP1985))
```

とすれば交互作用項の係数の有意性をみることができて，有意確率が 0.118 という結果が得られるので傾きには差がないとわかる。最後に

```
summary(glm(CAR1990 ~ as.factor(REGION)+DIDP1985))
```

として as.factor(REGION)2 の有意確率をみると 0.05 より遥かに小さいので，修正平均にも差があるとわかる。つまり，東日本と西日本では，人口集中地区への居住割合の影響を調整しても，世帯当たりの自動車保有台数には有意に差があるといえる。

13.5 補足：一般線型混合モデル

複数の対象についての経時的観察データが複数あるときに，個体間の経時的な変化のパタンの違いをモデルに取り込むことによって一般化線型モデルをさらに一般化したのが一般線型混合モデル (General Linear Mixed Model) である。高度な分析なのでここでは説明しないが，非常に強力である。R では，nlme というライブラリが提供されている。8 歳から 14 歳まで 2 年おきに歯列矯正の指標として，頭蓋の X 線写真により下垂体から翼上顎裂までの距離 (mm) を，男児 16 人，女児 11 人について測定したデータ (Orthodont という組み込みデータ) による実行例は，library(nlme) としてから，example(lme) とすれば見ることができる。年齢によるモデル，性と年齢と個体差によるモデルについて出力される。