第 6 章　線形回帰モデル

２変量 X と Y の関係を分析するために、Y を X の一次式の形でとらえる方法があります。これを線形単回帰といいます。ここで、X は説明変数、Y は被説明変数です。

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

は回帰直線、α，β は回帰係数です。これらの回帰係数は固定された母数ですが、決して既知ではありません。また、X は非確率変数です。

$$E(Y|X_i) = \alpha + \beta X_i$$

となり、$E(|)$ は条件付き平均です。$E(|)$ は確定的な、またはシステマティックな部分です。これを母集団線形回帰モデルといいます。

また、

$$Y_i - E(Y|X_i) = \epsilon_i$$

と書き、$\hat{\epsilon}_i$ は攪乱項です。これは確率的または非システマティックな部分を表現しています。

6.1 確率的誤差項のもつ意味

ϵ_i はこのモデルでは説明できない部分です。ではなぜ ϵ_i が必要なのでしょうか。主に 7 つの理由があります。

- 理論のあいまいさ

理論がまだ確立されていない。または理論で説明できない部分が残っている状態です。

- データの不完全性

十分なデータ、情報が与えられていない場合です。

- 主変数とそれ以外の変数

いくつかの変数による影響が確率的な効果として表れています。

- 内在的な確率的要素
- データに含まれるノイズ

データにはノイズが含まれていてそのノイズを誤差項で表現しています。

- 単純性の原理

回帰モデルを単純な形にしておきたい場合などです。

- 間違った仕様

正しくない関数形式を用いている場合などです。

6.2 線形単回帰モデル

母集団は一般に手に入りません。したがって、母数も未知です。そこで、母集団回帰式と同様に標本回帰式を導入します。それを

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

と書き、$\hat{Y}_i$は$E(Y|X_i)$の推定量、$\hat{\alpha}$，$\hat{\beta}$は α，β の推定量です。これらは統計量です。統計量とは、与えられた標本データから得られる情報を用いて母数を推定する方法、関数です。推定量は確率変数ですが、推定値は実際のデータから計算された値です。

推測統計量を用いると

$$Y_i = \hat{\alpha} + \hat{\beta} X_i + \hat{\epsilon}_i$$

と書けます。$\hat{\epsilon}_i$を回帰残差項といいます。これは線形単回帰モデルです。$\hat{\epsilon}_i$は互いに独立に正規分布$N(0, \sigma^2)$にしたがいます。

6.3 最小二乗法

母集団線形回帰関数は

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

です。また、標本線形回帰関数は

$$\begin{aligned} Y_i &= \hat{\alpha} + \hat{\beta} X_i + \hat{\epsilon}_i \\ &= \hat{Y}_i + \hat{\epsilon}_i \end{aligned}$$

です。したがって、

$$\begin{aligned} \hat{\epsilon}_i &= Y_i - \hat{Y}_i \\ &= Y_i - \hat{\alpha} - \hat{\beta} X_i \end{aligned}$$

となり、$\hat{\epsilon}_i$の平方和は

$$\begin{aligned} \sum \hat{\epsilon}_i^2 &= \sum (Y_i - \hat{Y}_i)^2 \\ &= \sum (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2 \end{aligned}$$

となります。α と β は、これを最小になるように推定して得ます。すると、

$$\hat{\beta} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

が得られます。

このようにして求めた統計量を最小二乗統計量といいます。

- これらの統計量は標本から計算されていて簡単に手に入ります。

- これらは点推定量です。

- 標本回帰直線が得られます。

  - 標本回帰直線は Y_i と X_i の平均を通ります。
  - $\bar{\hat{Y}} = \bar{Y}_i$ です。
  - $\hat{\epsilon}_i$=0 ですので、$Y_i - \bar{Y} = \beta(X_i - \bar{X})$が得られます。$\bar{X}$ は X_i の標本平均です。

6.3.1 古典的最小二乗法の仮定

統計モデルは幾つかの仮定のもとに成り立ちます。線形単回帰モデルの仮定をここで列挙します。

回帰関数は線形でなければならない。
X_i は確率変数であってはならない。確率変数の場合には誤差項と独立でなければならない。
ϵ_i の平均はゼロである。
ϵ_i の分散は一定である。
ϵ_i とϵ_j の共分散はゼロである。
観測値の数 n は説明変数の数よりも多い。
X_i は一定であってはならないが、外れ値などがあってもならない。

6.3.2 推定の信頼性

単純な標本回帰直線には、$\hat{\alpha}$, $\hat{\beta}$, $\hat{\epsilon}_i$の 3 つの推定量があります。これがどの程度信頼できるのかを見ていきます。

標準誤差

標準誤差(Standard Error,Std Err)は統計量のばらつきの度合いを示す尺度です。回帰係数 α, β の推定値の標準誤差は

$$se(\hat{\alpha}) = \sqrt{\frac{\sum X_n^2}{N\sum(X_n - \bar{X})^2}}\sigma$$

$$se(\hat{\beta}) = \frac{\sigma}{\sqrt{\sum(X_n - \bar{X})^2}}$$

で与えられます。ここで σ は、ϵ_iの分散で推定の標準誤差と呼ばれ次式で与えられます。

$$\sigma^2 = \sum \frac{\epsilon_i^2}{N} = \frac{\sum([Y_n - E(Y|X_n)]^2}{N}$$

この母数は未知ですので、その推定量は

$$\hat{\sigma}^2 = \sum \frac{\hat{\epsilon}_i^{\,2}}{N-2} = \frac{\sum([Y_n - E(Y|X_n)]^2}{N-2}$$

です。$E(\hat{\sigma}^2) = \sigma^2$ です。標準誤差は、統計量のバラツキ具合の尺度であり、標準誤差の推定値は標本回帰直線が観測値をどの程度説明できているかを示す適合度(goodness of fit)の目安となります。

決定係数

R^2(R-squared)は、標本回帰直線がデータをどの程度説明しているかを示す指標、適合度(goodness of fit)を表す測度として知られています。それは、決定係数(coefficient of determination)と呼ばれ、次式で与えられます。

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{\mathrm{ESS}}{\mathrm{TSS}}$$

ここで$\bar{Y}$は Y の標本平均です。決定係数が 1 に近いほど、相対的なバラツキは少なくなります。ESS(Explained Sum of Squares, 回帰平方和)、TSS(Total Sum of Squares, 全体平方和)です。

動調整済み決定係数(Adj R-squared)

これは説明変数の数の効果を考慮した係数です。説明変数の数が多くなると決定係数は良くなる傾向があります。したがって、その分を調整します。

回帰係数の区間推定と仮説検定

- H_0:μ=0, H_1: μ≠ 0 とします。

回帰係数の母平均をゼロとし、標本平均を $\hat{\theta}$ とすると有意水準 α の信頼区間は

$$0 - t_{\alpha/2} se(\hat{\theta}) \leq \hat{\theta} \leq 0 + t_{\alpha/2} se(\hat{\theta})$$

となります。これを採択域(the region of acceptance)といい、この外側の領域を棄却域(the reion(s) of rejection)といいます。これを危険域(the critical region(s))と呼ぶことがあります。大きな$|t|$値は帰無仮説の棄却域にいることになります。また、

$$t = \frac{\hat{\theta}}{se(\hat{\theta})}$$

からp値(p-value)を計算できます。

一般に、

関係 (p は p 値を表す)	解釈
$0.01 \geq p$	帰無仮説を棄却する。
$0.1 \geq p \geq 0.01$	帰無仮説を棄却するに足る。
$p \geq 0.1$	帰無仮説を棄却するのは難しい。

と解釈されます。

「帰無仮説を棄却する」とは0.01以下の確率でしか起こらないことが起こった、ということです。
「帰無仮説の棄却は難しい」は棄却するに十分な証拠がないということです。
このように表現する理由として、統計学の目的は極力誤った判断を減らすことにあるからです。

分散分析

Y_iの平均と$\hat{Y}$の平均が同じ母集団から得られたのかどうかを検定します。Y_iの母平均(μ)と$\hat{Y}$の母平均(μ_e)の差が有意であるかどうかを調べます。Y_iと$\hat{Y}$をそれぞれ群としてとらえます。そしてこの2つの群のそれぞれの平均に違いがあるかどうかを調べます。これらの標本の群の間で平均値に差があるからといって、母平均にも差があるとは限りません。Yとその予測値の2群の母平均に違いがなくても無作為に抽出された標本には違いがあるのかもしれません。このようなときに分散分析が用いられます。分散分析と名前がついていますが、実は平均値の差の検定です。

分散分析では非説明変数の全体平方和(TSS; 分散×自由度)を回帰モデルで説明できる部分とできない部分に分解します。説明できる部分が回帰平方和(ESS)、できない部分が残差平方和(RSS)です。したがって、TSS=ESS+RSS です。

- $H_0 : \mu_t = \mu_e$
- $H_1 : \mu_t \neq \mu_e$
  - μ_t: 被説明変数の母平均
  - μ_e: 予測値の母平均

$$\begin{aligned} \text{TSS 全体平方和} &= \sum(Y_i - \bar{Y})^2 \\ \text{ESS 回帰平方和} &= \sum(\hat{Y}_i - \bar{Y})^2 \\ \text{RSS 残差平方和} &= \sum(\hat{Y}_i - Y_i)^2 \\ \text{ESS} &= \text{TSS-RSS} \end{aligned}$$

ここで、回帰モデルで説明できる部分とできない部分を自由度て調整して

$$F = \frac{\text{ESS}/df_p}{\text{RSS}/df_r}$$

とします。df_p は群間の自由度です。この場合は 2−1=1 です。df_r は群内の自由度です。この場合は郡内のデータ数−２です。回帰分析では RSS が小さいと推定モデルは Y_i をよくとらえていることになります。つまり、回帰係数に意味があるということです。そうすると ESS は大きくなり TSS に近づきます。回帰係数に意味がなければ、RSS は TSS に近づきます。その際には ESS は小さくなります。それは Y_i の平均と $\hat{Y}_i$ の平均が同じ母集団から得られたということです。したがって、帰無仮説はすべての回帰係数がゼロであると書き換えることができます。R^2=ESS/TSS を用いると

$$F = \frac{R^2}{(1-R^2)}\frac{df_r}{df_p}$$

と変形できます。

$F \geqq F_\alpha(df_p, df_r)$ ならば帰無仮説を棄却します。
$F < F_\alpha(df_p, df_r)$ ならば帰無仮説を棄却しません。
また、p−値を用いることもできます。

例　年齢と性別の同じ人の体重と身長の関係が与えられたとして、エクセル分析ツールの回帰分析を用いて分析します。結果はつぎのようになります。

	A	B	C	D	E	F	G	H	I	J
1	身長	体重		概要						
2	113	20								
3	111	19		回帰統計						
4	118	20		重相関 R	0.65					
5	112	16		重決定 R2	0.43					
6	114	19		補正 R2	0.39					
7	118	23		標準誤差	2.59					
8	124	22		観測数	19					
9	118	20								
10	122	23		分散分析表						
11	119	19			自由度	変動	分散	観測された分散比	有意 F	
12	118	22		回帰	1	85.48	85.48	12.69	0.00	
13	116	22		残差	17	114.47	6.73			
14	118	21		合計	18	199.95				
15	114	19								
16	119	22			係数	標準誤差	t	P-値	下限 95%	上限 95%
17	117	20		切片	93.52	6.72	13.91	0.00	79.34	107.71
18	119	20		X 値 1	1.15	0.32	3.56	0.00	0.47	1.84
19	119	24								
20	120	20								

回帰統計を見ると、決定係数 R2 と補正 R2 はそれぞれ 0.4 と 0.36 程度です。特に大きいとはいえません。

分散分析表の有意 F は F 検定の p–値のことです。小さいので、Y の平均とその推定値の平均が同じ母集団であるという帰無仮説を棄却します。つまり、回帰係数は有意であるという結論です。その下の表はそれぞれの回帰係数がゼロであるかないかの仮説検定をしています。P–値はゼロと判断できる程度に小さいので、帰無仮説は棄却され、回帰係数は有意であることが分かります。

6.4 線形回帰の意味

回帰直線が理解できたところで、線形回帰の意味を理解しましょう。いままで回帰直線を中心に線形回帰を説明してきました。したがって、線形回帰はデータを直線でとらえると理解されているのではないでしょうか。しかし、実際は線形回帰の’線形’はデータを線形に結合するという意味なので、幅広い分析に使えます。たとえば、

$$Y=a+bX^2$$

でも

$$Y=a+bX^3$$

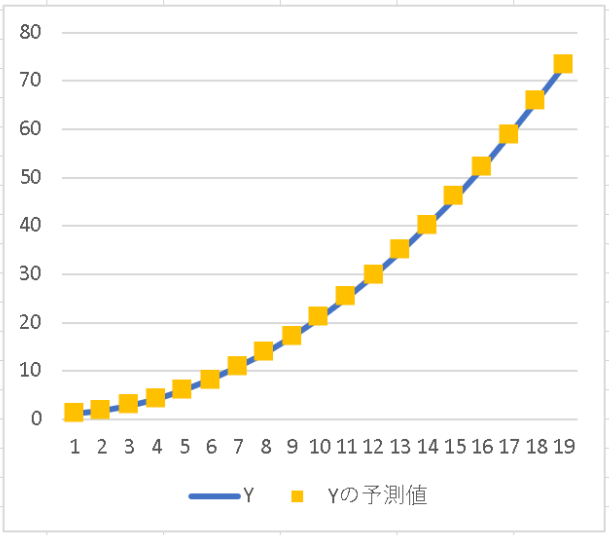
出も構いませんし、

$$Y=a+b\sin(\omega t)$$

でも構いません。

例　X を 1～19 までの整数として $Y = 0.2\cdot X^2 + 1$ を算出し、Y を被説明変数、X^2 を説明変数としてエクセルの LINEST 関数を用いて線形回帰分析を行います。a=1，b=0.2 と算出されます。グラフは青線が Y、■が予測値です。

	A	B	C	D	E	F	G	H	I	J
1	Y	X	X^2	Yの予測値						
2	1.2	1	1	1.2						
3	1.8	2	4	1.8						
4	2.8	3	9	2.8						
5	4.2	4	16	4.2						
6	6	5	25	6						
7	8.2	6	36	8.2						
8	10.8	7	49	10.8						
9	13.8	8	64	13.8						
10	17.2	9	81	17.2						
11	21	10	100	21						
12	25.2	11	121	25.2						
13	29.8	12	144	29.8						
14	34.8	13	169	34.8						
15	40.2	14	196	40.2						
16	46	15	225	46						
17	52.2	16	256	52.2						
18	58.8	17	289	58.8						
19	65.8	18	324	65.8						
20	73.2	19	361	73.2						
21										
22		a=	1	=INDEX(LINEST(A2:A20,C2:C20),2)						
23		b=	0.2	=INDEX(LINEST(A2:A20,C2:C20),1)						



6.5 重回帰分析

分析をする際には、厳密に周囲の環境を一定に保ち、変更は１つに絞るべきであるという考え方があります。単回帰分析を信奉しすぎています。１つだけ変更するとしばしばうまくいかないこともあります。注意深く考え抜かれた計画にしたがえば本来の性質に近づけるかもしれません。線形回帰分析の中で説明変数(要素)の数が 2 つ以上のものを重回帰分析といいます。被説明変数(目的変数)が複数の要素により影響を受けているときなどには有効な分析手法となります。また、回帰誤差の減少につながる可能もがあります。

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_m X_{mi} + \varepsilon_i, i = 1, 2, \cdots, n$$

単回帰分析の７つの仮定に加えて、変数 X の一部またはすべての間で無相関であり、つまり多重共線性があってはなりません。またモデルの仕様は適切である必要があります。それぞれの説明変数 X_{mi}、または要素を構成する一点一点のデータの分散が不均一である分散不均一性や、その間に相関関係のある自己相関などがあってはなりません。モデルを過剰適合(オーバーフィッティング)したり、関連のあるデータを削除してしまったりする結果になりかねません。

説明変数が複数あると、桁数が大きく違ったり、分散の大きさに違いがあったりします。それでも、最小二乗法は適切な回帰係数を算出してくれます。これを編回帰係数といいます。しかし、編回帰係数は目的変数への影響度を示しているわけではありません。各要素の目的変数への影響度合いを知りたい場合には、それを調整する必要があります。編回帰係数に説明変数の標準偏差を掛け、目的変数の標準偏差で割ることで調整できます。

6.5.1 多重共線性

説明変数 X_m の間に強い相関があってはなりません。そのような特性を多重共線性といいます。これにより OLS による推定値の分散や共分散が大きくなったりします。したがって、

- 区間推定の幅が広くなり、帰無仮説を受け入れやすくなります。
- それぞれの回帰係数の推定値の p-値が統計的に有意ではなくなります。
- R^2 が大きく見積もられます。
- 回帰係数の推定値や標準誤差が説明変数の小さな変化に敏感になります。

したがって、多重共線性は避けなければなりません。その際の説明変数の選択にはさまざまな方法がありますが、基本は残差のもつ特性を把握しつつ、多くの組み合わせを試してみることです。

例　スペクターとマッテーオ個人向け教育プログラムによる効果の実データ

変数名	説明
Grade	生徒の評価点(GPA)が改善したかどうかの２値データ。１が改善。
TUCE	経済のテストの評価点。
GPA	生徒の評価点(GPA)の平均。
PSI	プログラムへの参加の可否。

についてエクセル分析ツール(回帰分析)を用いて重回帰分析を行います。

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	成績の改善	経済の成績	生徒の評価点	参加・不参加	残差	概要							
2	0	2.7	20	0	0.06								
3	0	2.9	22	0	−0.06	回帰統計							
4	0	3.3	24	0	−0.27	重相関 R	0.68						
5	0	2.9	12	0	0.03	重決定 R2	0.46						
6	1	4.0	21	0	0.44	補正 R2	0.40						
7	0	2.9	17	0	0.00	標準誤差	0.37						
8	0	2.8	17	0	0.05	観測数	32						
9	0	2.9	21	0	−0.04								
10	0	3.0	25	0	−0.16	分散分析表							
11	1	3.9	29	0	0.39		自由度	変動	分散	観測された分散比	有意 F		
12	0	2.6	20	0	0.08	回帰	3	3.32	1.11	7.93	0.00		
13	0	3.3	23	0	−0.27	残差	28	3.90	0.14				
14	0	3.6	23	0	−0.39	合計	31	7.22					
15	1	3.3	25	0	0.73								
16	0	3.5	26	0	−0.40		係数	標準誤差	t	P-値	下限 95%	上限 95%	調整係数
17	0	2.7	19	0	0.04	切片	−1.51	0.50	−2.99	0.01	−2.54	−0.47	−1.46
18	0	2.8	25	0	−0.03	経済の成績	0.46	0.16	2.96	0.01	0.14	0.78	3.72
19	0	2.8	19	0	0.00	生徒の評価点	0.01	0.02	0.58	0.57	−0.03	0.05	0.01
20	0	3.1	23	0	−0.18	参加・不参加	0.43	0.13	3.20	0.00	0.16	0.71	0.32
21	1	3.2	25	1	0.35								
22	0	2.1	22	1	−0.11								
23	1	3.6	28	1	0.10								
24	0	2.9	14	1	−0.41								
25	0	3.5	26	1	−0.83								
26	1	3.5	24	1	0.18								
27	1	2.8	27	1	0.48								
28	1	3.4	17	1	0.33								
29	0	2.7	24	1	−0.42								
30	1	3.7	21	1	0.16								
31	1	4.0	23	1	−0.02								
32	0	3.1	21	1	−0.58								
33	1	2.4	19	1	0.77								
34	0.48	0.47	3.90	0.50	0.35								

- 回帰統計の結果から決定係数が 0.46 で、被説明変数の 46%が説明変数で説明されます。
- 分散分析表からは有意 F が 0.00 で、すべての回帰係数がゼロという帰無仮説は棄却されました。
- 回帰係数の分析では X 値 2 の P-値が 0.57 と大きく、多重共線性が疑われます。プログラムの参加の可否と経済の

成績が Grade の改善に寄与しているようです。

- 調整係数は説明変数の標準偏差(B34, C34, D34)と被説明変数(A34)の標準偏差で調整したもので、回帰係数の影響度合いを比較できます。

6.6 尤度

標本から母数を推定するために、点推定、区間推定を行うと同様に、ある分布を仮定して、その分布が再現される確率が最も高くなるようにパラメータを決めるという方法があります。パラメータを θ とすると

$$L(\theta|x)=f(x|\theta)$$

となる L を尤度関数といいます。これは母数の点推定です。尤度を最大にするのも、尤度の自然対数を最大にするのも同じなので、扱いやすい対数尤度を最大にします。

正規分布を仮定する場合には、母数は $\theta=(\mu, \sigma^2)$ なので最尤推定量は

$$\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2) = \left(\bar{x}, \frac{(\sum x_i - \bar{x})^2}{n} \right)$$

となります。尤度そのものではなくその比率に大きな意味があります。たとえば

$$L(\theta_1|x) / L(\theta_2|x) > 1$$

であれば、θ_1 の方が θ_2 よりももっともらしいと判断できます。

この性質を利用したものとして情報量基準があります。これは統計的モデルの構築の際に用いられます。情報量基準は限られたデータにもとづいて良いモデルを選択するための方法を示唆してくれます。情報量基準は有限なデータを利用してモデルを構築する際に、自由度の大きなモデルを用いることによる不安定性の上昇の度合いを避けるために、パラメータ数の決定や、目的変数の選択に利用されます。情報量基準を最小にすることで情報量基準という意味で最良のモデルを選択できます。主な情報量基準に赤池情報量基準(AIC)とベイズ情報量基準(BIC)があります。

$$AIC= - 2 \ln L +2n$$

$$BIC= - 2\ln L +2\ln(n)$$

で表わされます。AIC は主に予測に用いられ、BIC は検証に用いられます。

練習問題 6.1：身長・体重の単回帰分析の例の有意 F について、エクセルシートを使って自分で求めてみましょう。
練習問題 6.2: 赤ワインデータについて重回帰分析により評価を予測できるどうか確かめて見よ。
練習問題 6.3：AIC と BIC を用いてスペクターとマッテーオのデータのモデルを再度検討してみてください。