原稿作成日: 2019 年 3 月 29 日

線形回帰モデル

<教材提供>
AMED 支援「国際誌プロジェクト」 提供

目次

はじめに

１時点で採取した連続変数のアウトカムに対しては、**線形回帰モデル**が用いられ、このようなモデルが用いられる解析は線形回帰分析と呼ばれます。この単元では線形回帰モデルの使い方と解析結果の解釈の仕方を説明します。また線形回帰分析を行う際に注意すべき仮定や、その確認方法、またその仮定が満たされていない場合の対策方法まで解説します。

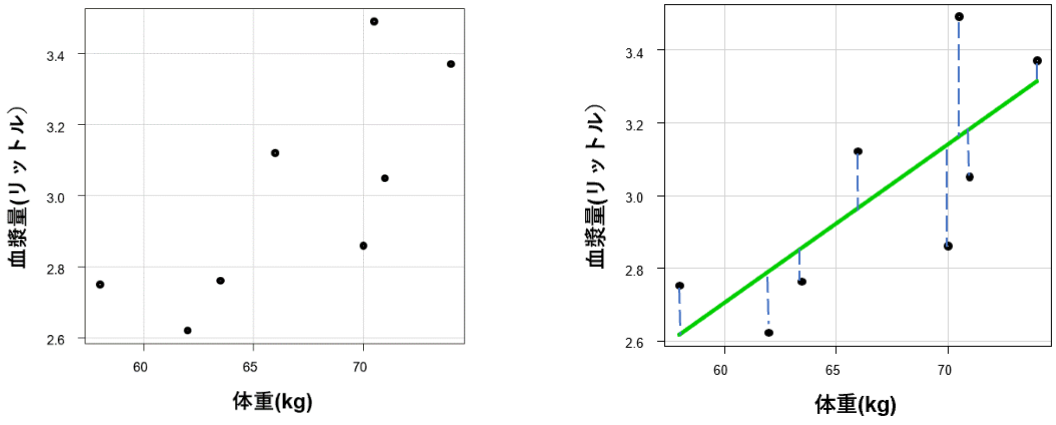
学習目標

本単元を通じてあなたが習得を目指すのは

- 線形回帰分析とは何か、どのような状況で利用可能かを理解する
- 線形回帰分析の仮定を理解しその確認・対策方法を習得する
- 線形回帰分析の結果の解釈・報告を適切にできるようになる

線形回帰モデルのあてはめと最小二乗法

以下のデータは 8 人の患者さんから血漿量と体重のデータを集め、その関連を散布図により表したものです。このデータからは体重とともに血漿量が増加している傾向が見られています。この傾向をとらえるために一本の直線を引くとしたらどのような方法をとるべきでしょうか。



この直線を引くための方法の一つとして「**最小二乗法**」という方法があります。最小二乗法では、上右図のように、全データ点の真ん中を通るように直線を引きます。より厳密には、各データ点からＹ軸と平行にこの直線までの距離（多変量解析の単元でも説明した通り、この距離を「残差」と呼びます）を求め、全データ点にわたって、その距離を二乗したものの和（この和を残差平方和と呼びます）をとり、それを最小にするような直線を求めます。

なぜ二乗するかというと、直線の上に並んだデータまでの距離はプラスとなり、直線の下に位置するデータまでの距離はマイナスであるため、二乗しないでそのまま全てを合計した値は０になってしまうからです。これを

避けるために、全ての残差を二乗してマイナスをプラスに変えた後にその合計を計算し、合計値が最小になる位置に直線を引きます。このようにして得られる直線は、**線形回帰直線**と呼ばれます。上述の計算方法では、残差平方和(残差の二乗の和)を最小にしているので、最小二乗直線又は最小二乗回帰直線とも呼ばれます。

それでは EZR を用いて、血漿量(変数名:Plasma)を被説明変数(y)とし、体重(変数名:Weight)を説明変数として、最小二乗直線を求めてみましょう。作成の仕方はビデオをご覧ください。

下の表が EZR での計算結果です。

回帰モデル①

	回帰係数推定値	標準誤差	t 統計量	P 値
(Intercept)	0.086	1.024	0.084	0.936
Weight	0.044	0.015	2.857	0.029

(Intercept)の横にある回帰係数推定値は回帰直線の切片、Weight の横にある回帰係数推定値は回帰直線の傾きを示しています。P 値はこれらの切片や傾きが 0 かどうかを検定しています。有意水準として両側 5%を用いた場合、この例では傾き 0.044 は統計的に 0 とは異なる(体重と血漿量の間には関連がある)ことが分かりました。

この直線は

$$\hat{y} = 0.086 + 0.04x$$

という式で表すことができます。左辺(y)には臨床研究で用いられるアウトカムにあたる被説明変数の予測値が入ります。また左辺の被説明変数の予測値(y の予測値)は、手元のデータ(標本)を推定した線形回帰モデルにあてはめて予測された値であるため、y に^(ハット)を添えて$\hat{y}$と記載しています。統計学では予測値や推定値などを区別する時にはハットを用いて表現することを覚えておいてください。一方、右辺(x)には効果を検証したい暴露因子や、背景因子など説明変数が入ります。0.086 という数字は、先ほど説明したように回帰直線の切片を表しています。切片とは、xが 0 の時の期待されるyの値ということなので、体重 0kg の人の予想される血漿量と理解できます。また(Intercept)の行にある P 値は 0.936 で、これは体重 0 の人の血漿量が 0 であるかどうかの検定を行っています。P 値が 0.05 より大きければ、切片の値は 0 とは区別できない、0.05 より小さければ切片の値は 0 とは統計的に異なるという結論になります。しかしそもそも体重が 0kg の人は存在しないので、回帰モデルの結果においては、傾きの値に着目されることが多く、切片の値のみに着目することはあまりありません。

一方傾きが 0 であれば、最小二乗直線は*x*軸と並行なグラフになってしまうので、この場合は体重がいくら増減しても、予想される血漿量は変わらないということになります。つまり 2 つの変数の間には関連がないということになります。よって 2 つの変数の関連を検討する場合、最小 2 乗直線の傾きが 0 かどうかに着目することになります。

最小二乗直線は*x*によって*y*を予測する場合にも用いることができます。すなわち、この最小二乗直線を用いれば、患者さんの体重を測定しさえすれば、その患者さんの血漿量を予測することができます。例えばこの最小二乗直線の式にあてはめると、体重 60kg の患者さんの血漿量の予測値は、

$$0.086 + 0.04 \times 60 = 2.486$$

と計算できます。関連を検討するときには重要でなかった切片の値も、説明変数の値が与えられたときにアウトカムの予測を行うときには重要な役割を果たします。

多変量線形回帰モデルを予測に用いる

1 つの*y*(被説明変数)を複数の*x*(説明変数)で予測する場合、多変量線形回帰モデルを用いることができます。例えば、慢性腎臓病の患者さんと健常者を対象として、炎症マーカーである CRP を測定し、慢性腎臓病(変数名:ckd_yes ; 1＝慢性腎臓病、0＝健常者)であるか否か、年齢(変数名:age)、BMI(変数名:bmi)、喫煙の有無(変数名:smoke_yes ; 1＝喫煙あり、0＝喫煙なし)の 4 つの変数が炎症に関連しているかを調べてみました[1]。EZR を用いて多変量線形回帰モデルをあてはめた結果は以下のようになります。(実際の解析方法はビデオで紹介します)

多変量線形回帰モデルのあてはめ結果②

	回帰係数推定値	95%信頼区間下限	95%信頼区間上限	標準誤差	t統計量	P値
(Intercept)	-4.54812899	-11.56559956	2.46934157	3.56006450	-1.2775412	0.20280194
age	0.01899569	-0.05741352	0.09540489	0.03876350	0.4900406	0.62460953
bmi	0.17692193	0.04078492	0.31305894	0.06906428	2.5616995	0.01110692
ckd_yes	2.90555549	0.36923393	5.44187705	1.28671268	2.2581230	0.02495149
smoke_yes	0.65309934	-1.25723468	2.56343336	0.96914013	0.6738957	0.50110829

t 統計量とは、回帰係数推定値をその標準誤差で割ることで得られ、それによって P 値の値が決まります。95%の信頼区間は回帰係数推定値に約 2 倍した標準誤差を足し引きして得られます。上記のような解析結果を論文や学会で公表する際には、回帰係数推定値、信頼区間、P 値を記載すれば十分です。統計量は P 値と一意に関連するので、標準誤差ではなく 95%の信頼区間を、t 統計量ではなく P 値を示せば最低限伝えるべき情報は含まれています。

上記の解析結果の解釈は

❶年齢が 1 歳増えるごとに CRP の予測値は 0.019 増えるが、この上昇は統計的に有意ではない（P＝0.625）
❷BMI が 1 増えるごとに CRP の予測値は 0.177 増える。この上昇は統計的に有意である（P＝0.011）
❸慢性腎臓病の患者は健常者に比べ CRP の予測値は 2.906 高い。この差は統計的に有意である（P＝0.025）
❹喫煙者は非喫煙者に比べ CRP の予測値は 0.653 高いが、この差は統計的に有意ではない（P＝0.501）

また、多変量線形回帰モデルを用いた CRP の予測値は以下のようになります。

CRP の予測値 ＝ -4.548 ＋ 0.019 × 年齢 ＋ 0.177 × BMI ＋ 2.906 × 慢性腎臓病の有無 ＋ 0.653 × 喫煙の有無

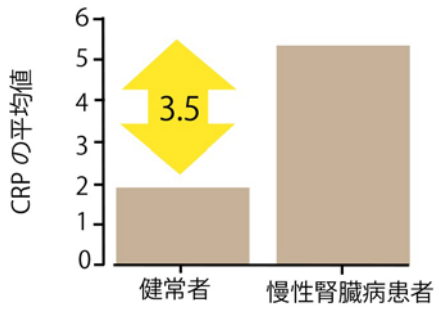
この式を用いて、年齢 60 歳、BMI が 25 で慢性腎臓病を患っていて、喫煙をしている人の CRP の予測値は

CRP の予測値 ＝ -4.548 ＋ 0.01 × **60** ＋ 0.177 × **25** ＋ 2.906 × **1** ＋ 0.653 × **1** ＝ 3.383

と計算できます。

この結果から年齢、BMI、喫煙の有無の違いを調整した慢性腎臓病患者と健常者の CRP 値の違いを検討することも可能です。例えば、慢性腎臓病患者の CRP の平均値は 5.40、健常者では 1.90、すなわち、群間差は 3.5 ですが、単純に Student の t 検定で群間比較を行うと、統計的に有意な差であることが確認されました（P＝0.005）。しかし、この群間差が単に慢性腎臓病患者と健常者の違いによるものかは、年齢、BMI、喫煙率等の背景を考慮しないと判断できません。

背景のズレを無視して CRP の平均値を比べた結果



そこで比較群間の背景を比較してみたところ、慢性腎臓病患者の方が、平均して 8 歳高齢であることが分かりました。

背景比較

	健常者 N＝39	慢性腎臓病患者 N＝180	P 値
喫煙割合	51.30%	36.70%	0.105
平均年齢	58.59	66.33	<0.001
平均 BMI	28.39	30.36	0.118

一般に年齢が高くなるとCRP が上昇しやすいということを考慮すると、慢性腎臓病患者と健常者の CRP の平均値の 3.5 の差は、慢性腎臓病が原因なのか、年齢が 8 歳高齢であることが原因なのかわからないことになります。このように背景の違いにより、比較したい慢性腎臓病と健常人の比較ができなくなることを交絡が起こっていると呼びます。

背景情報も説明変数として多変量線形回帰モデルに投入することで、この背景の差を統計的に調整したうえで、慢性腎臓病と健常人との比較が可能になります。このメカニズムを多変量線形回帰モデルの式を用いて説明します。

多変量線形回帰モデルのあてはめ結果より、

❶年齢が 1 歳増えるごとに CRP の予測値は 0.019 増える。
❷BMI が 1 増えるごとに予想される CRP の予測値は 0.177 増える。
❸喫煙者は非喫煙者に比べ CRP の予測値は 0.653 高い。

ことが分かります。

また、背景を比較した表より健常者に比べ慢性腎臓病患者は、平均年齢は 7.74 歳高く、平均 BMI は 1.97 高く、喫煙者の割合が 14.6%低いことが分かります。これを先の多変量線形回帰モデルの解析結果から得られる情報にあてはめると、慢性腎臓病患者は健常者に比べて CRP の予測値は、

$$0.019 \times 7.74 \ + \ 0.177 \times 1.97 \ + \ 0.653 \times 0.146 \ = \ 0.59$$

だけ過剰に高くなっていることがわかります。背景のズレを無視した解析によって得られた CRP 値の差である 3.5 から、上で計算した CRP 値の差である 0.59 を差し引くと、3.50 － 0.59 ＝ 2.91 となります。これは、多変量線形回帰モデルにおける慢性腎臓病の回帰係数の推定値、すなわち、多変量解析で背景を調整した慢性腎臓病の有無による CRP 予測値の差 2.91 と一致していることがわかります。このように、多変量線形回帰モデルを用いた解析では背景因子の影響も考慮して、背景のズレを統計的に調整（帳消し）にすることが可能です。

線形回帰モデルの成立条件

線形回帰モデルには、Student の t 検定と同様に、正しい解析結果を得るのに満たされるべき仮定があります。これらを以下にまとめます。

仮定その 1(観測値間の独立性)
　各研究対象者からの測定値はお互いに独立であること
　(データセットにおいて 1 対象者 1 行のデータが入力されているか否かで判断できます)

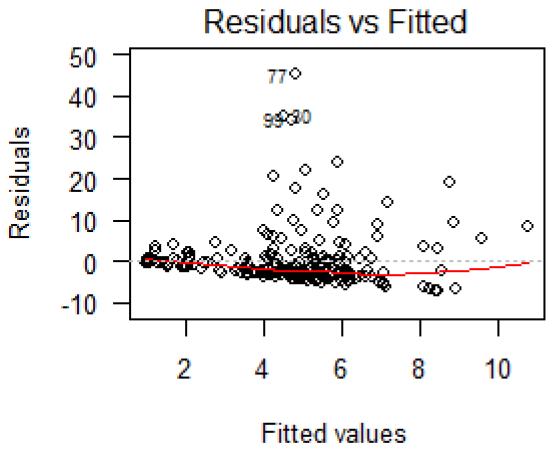
仮定その 2(誤差(又は被説明変数に関する観測値)の正規性)
　誤差(又は被説明変数に関する観測値)が正規分布に従っていること
　(線形回帰モデルをあてはめた後に得られる残差の分布を点検することで判断できます)

仮定その 3(誤差(又は被説明変数に関する観測値)の分散均一性)
　誤差(又は被説明変数に関する観測値)の分散が均一であること
　(線形回帰モデルをあてはめた後に得られる残差のバラツキが一様であることを点検することで判断できます)

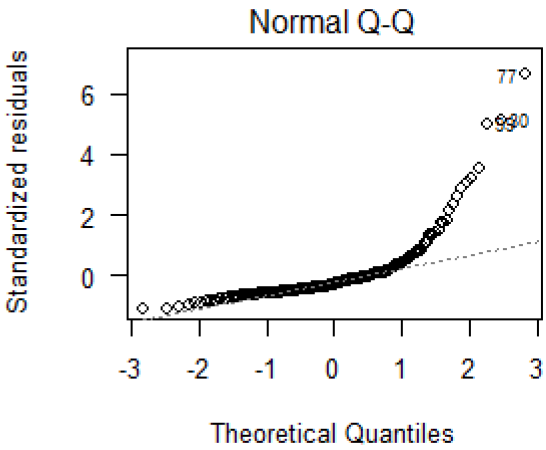
　誤差の正規性及び分散均一性は諸種の残差プロットというグラフから点検できます。実際のグラフの作成方法は、ビデオにて紹介します。

仮定が成り立っていない場合の残差プロット

（左）残差の分散が均一でない

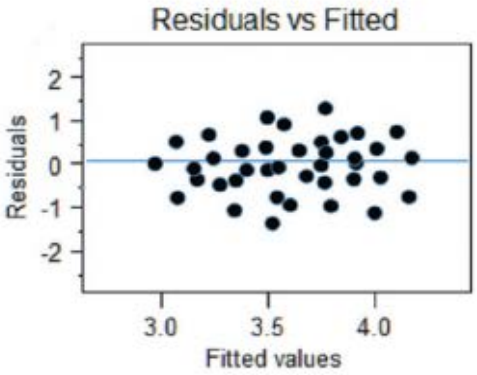


（右）残差が正規分布に従わない

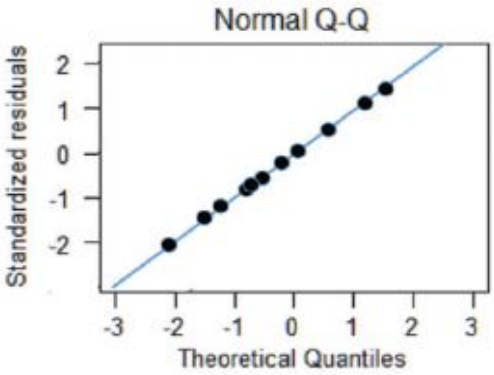


仮定が成り立っている場合の残差プロット

（左）残差の分散が均一である



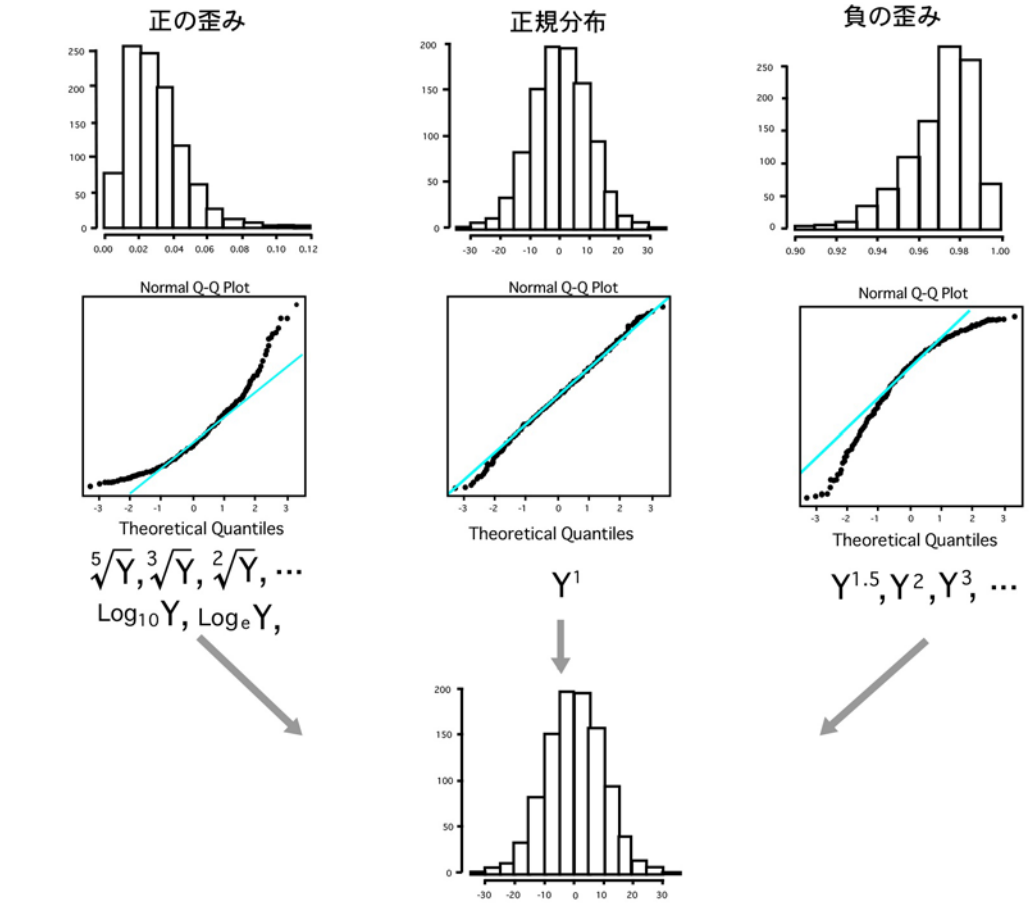
（右）残差の分布が正規分布に従う



残差の点検の結果、誤差が正規分布に従うことが疑われる場合、誤差（又は被説明変数に関する観測値）の正規性の仮定を満たすようにするには、被説明変数（アウトカム）の「変換（transformation）」が有効です。Normal Q-Q プロットが以下の図の左のパターンでは、残差の分布は正の歪みをもっており、このときには対数変換などを用いると、正規性の仮定を達成しやすくなります。反対に右のパターンでは、残差の分布は負の歪みをもっており、このときにはアウトカムの 2 乗、3 乗などの変換を用いると、正規性の仮定を達成しやすくなります。

線形回帰分析における被説明変数のデータ変換

AMED研究公正高度化モデル開発支援事業
国際誌プロジェクト



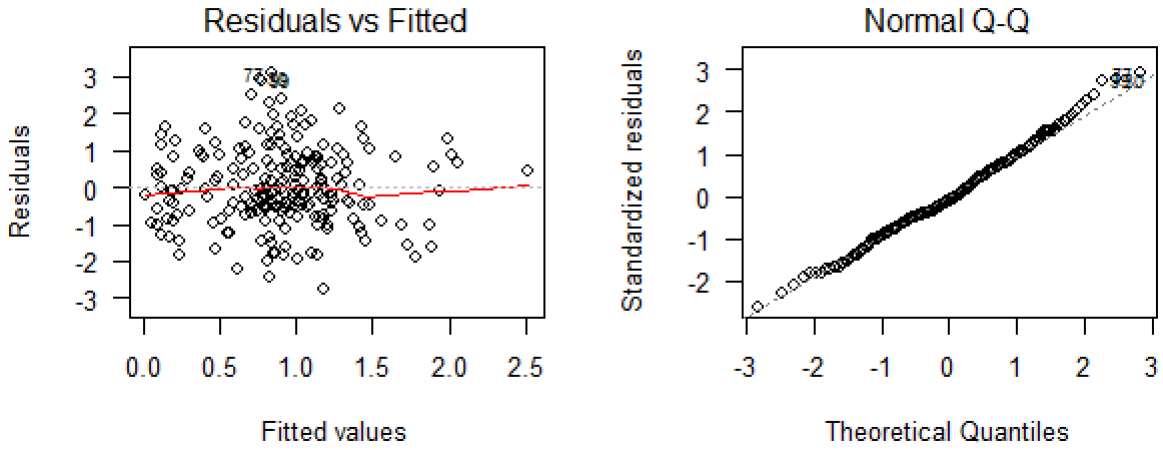
新谷歩「みんなの医療統計　多変量解析編 10 日間で基礎理論と EZR を完全マスター！」講談社[2] を参考に作成

変換を施した被説明変数を用いて線形回帰モデルによる解析を行い、残差の正規性を改めて点検します。ここで残差の正規性が確認されていれば、用いた線形回帰モデルにおける正規性の仮定は充足されていると考えられます。しかし対数変換を施したもとでの被説明変数を用いた線形回帰モデルは、結果の解釈が難しくなっ

てしまいます。そのため、同じ底を用いた指数変換によって対数を外すことで、結果の解釈を簡単にする作業が必要になります。この逆変換した結果は比較群間の差ではなく、比較群間の比で表現されます。それではこの一連の流れを以下で詳しく説明します。

自然対数で変換した CRP を用いた線形回帰モデルの結果

自然対数変換を施した CRP を被説明変数として線形回帰モデルをあてはめたところ、残差プロットは以下の図のようになりました。変換を施す前と比較すると、左図のように残差の分散もおおよそ均一になり、右図のように残差の正規性も達成されているようです。



以下のようなあてはめ結果が得られていますが、これは被説明変数に自然対数変換を施した状態（正確には尺度（scale）と呼びます）での結果であることに注意する必要があります。変換を施さない状態（尺度）でこの結果を解釈するには、ネイピア数（e）を回帰係数推定値で累乗する必要があります。因みに、被説明変数に常用対数変換（10 を底とした対数変換）を施した場合には、10 を回帰係数推定値で累乗する必要があります。

```
> multireg.table
             回帰係数推定値 95%信頼区間下限 95%信頼区間上限     標準誤差    t統計量          P値
(Intercept)   -1.506631221    -2.597755690    -0.41550675 0.553543253 -2.7217949 0.00703024700
age            0.006972178    -0.004908449     0.01885280 0.006027214  1.1567828 0.24865680584
bmi            0.046750220     0.025582704     0.06791774 0.010738588  4.3534794 0.00002080435
ckd_yes        0.638472262     0.244107585     1.03283694 0.200066915  3.1912936 0.00163036766
smoke_yes      0.025564750    -0.271467090     0.32259659 0.150688556  0.1696529 0.86544413859
```

被説明変数に自然対数変換を施して線形回帰分析を行った上記の結果は、次のように変換して解釈することができます。

❶年齢が 1 歳増えるごとに CRP の予測値は $e^{0.0069}$＝1.007 倍に増える。この上昇は統計的に有意ではない（P＝0.249）
❷BMI が 1 増えるごとに CRP の予測値は $e^{0.046}$＝1.047 倍に増える。この上昇は統計的に有意である（P＝0.00002）
❸慢性腎臓病の患者は健常者に比べ CRP の予測値は $e^{0.638}$＝1.892 倍である。この差は統計的に有意である（P＝0.0016）
❹喫煙者は非喫煙者に比べ CRP の予測値は $e^{0.026}$＝1.026 倍であるが、この差は統計的に有意ではない（P＝0.865）

線形回帰モデルを使用する際の注意点

繰り返しのないデータのみに使用可能

線形回帰モデルは、1 人の研究対象者から得られたアウトカムデータ（被説明変数に指定する変数）が 1 回のみ観測されている時に使用可能です。観測値が 1 人の研究対象者から繰り返し測定されている場合は利用できず、その場合は線形混合効果モデルや一般化推定方程式など、1 人の研究対象者からデータが繰り返し収集されていることを考慮にいれることのできる回帰モデルを使用する必要があります。

線形回帰モデルに入れられる説明変数の数

線形回帰モデルには、一つの目安として、症例数を 15 で割った数まで説明変数を入れることが考えられます（多変量解析の単元を参照）。例えば、150 人のデータがある場合、モデルに同時に入れることのできる説明変数は 10 個程度と考えられます。ただし、変数が 3 値以上のカテゴリー変数の場合は、カテゴリーの数から 1 を引いた数と考えます。例えば、5 つのカテゴリーの変数の場合は、4 個分の変数と考えます。

欠損値について

多変量線形回帰モデルに用いることのできるデータは、モデルで考慮した変数全てにおいて欠損がないデータのみです。逆にいえば、ある研究対象者のデータのいずれかの変数において欠損値がある場合には、当該研究対象者の欠損値のない変数のデータについても解析に用いられなくなります。そのため、欠損値の多い変数は、多変量線形回帰モデルに入れることを差し控えることも考えねばなりません。

この単元に関係する国際誌におけるチェックポイント： Annals of internal medicine のチェックリストなどに該当なし

本単元は、日本医療研究開発機構（AMED）が実施する研究公正高度化モデル開発支援事業（第一期）の「医系国際誌が規範とする研究の信頼性にかかる倫理教育プログラム」（略称：AMED 支援国際誌プロジェクト、信州大学・大阪市立大学）によって作成された教材です。作成および査読等に参加した専門家の方々の氏名は、こちらに掲載されています。

参考文献

[1] Luis F. Ramos, Ayumi Shintani, T. Alp Ikizler, and Jonathan Himmelfarb. Oxidative Stress and Inflammation Are Associated with Adiposity in Moderate to Severe CKD. J Am Soc Nephrol. 2008 Mar; 19(3): 593–599.

[2] 新谷歩. みんなの医療統計　多変量解析編　10 日間で基礎理論と EZR を完全マスター！. 講談社, 2017 年.