演習問題

$x_1, x_2 \ldots, x_n$ と $y_1, y_2 \ldots, y_n$ という n 個のデータに対し、$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ という回帰モデルを母集団に想定し、これらの係数に関して統計的推論を行う。これを踏まえて以下の問いに答えなさい。

基礎問題

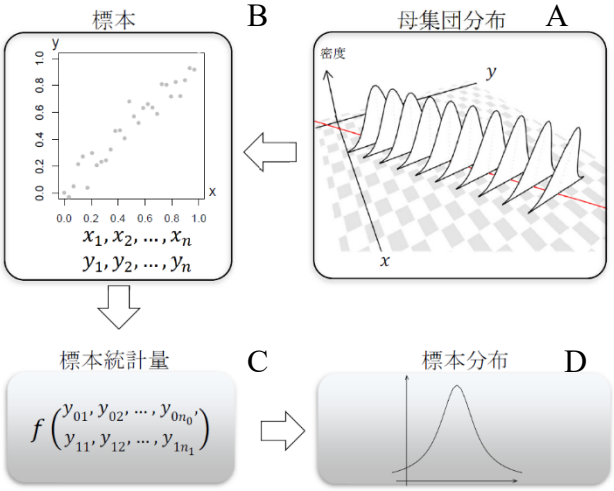
1 記法

問1 次の記法が何を表すか、日本語で説明しなさい。

(1) β_0　(2) β_1　(3) ε_i
(4) $y_i - \hat{y}_i$　(5) $\bar{y} - \hat{\beta}_1 \bar{x}$　(6) r_{xy}
(7) $\frac{s_{xy}}{{s_x}^2}$　(8) $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$　(9) $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

問2 下の図は、単回帰分析の説明としてこの授業で紹介された四つの概念の関係を表している。次の各量は A から D のどこにかかわる量であるか、最も適切なもの一つ選び記号で答えなさい。

(1) β_0
(2) $\hat{\beta}_0$
(3) $\beta_0 + \beta_1 x_i$
(4) ε_i
(5) e_i
(6) σ^2
(7) $\hat{y}_i$
(8) y_i
(9) s_{xy}/s_x^2
(10) $E[\hat{\sigma}^2]$
(11) r_{xe}
(12) $r_{y\hat{y}}$
(13) R^2
(14) h_i



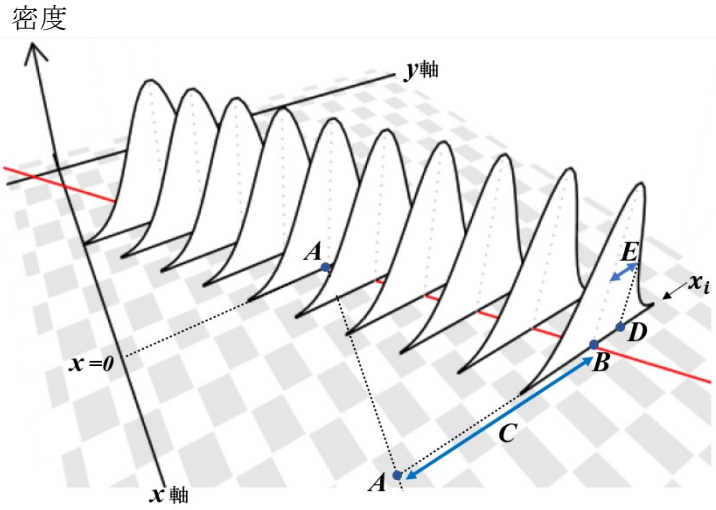
問3 次の記法で書かれた概念に対して続く日本語の説明が正しければ○と、間違っていれば×と記しなさい。また、誤りがある部分は適切に訂正しなさい。なお、そもそもそのような記法が存在しない場合には「該当概念なし」と記しなさい。

(1) s_y：母集団分布に想定される標準偏差。
(2) $\hat{s}_y$：母集団分布に想定される標準偏差の推定値。
(3) $\hat{\sigma}_{\bar{y}}$：母集団分布に想定される平均の標準偏差の推定値。

(4) $\hat{\sigma}_{\hat{y}}$：yの平均が従う標本分布の標準偏差の推定値。
(5) $\sigma_{\hat{\beta}_1}$：母集団に想定された直線の傾きを標本に基づいて推定したとき、その推定量がどのくらいばらつくのかを人間が推定したもの。

2　母集団における想定

下の図は、単回帰モデルに想定される母集団の構造を表した図である。A、B、C、D、E それぞれが何を表すか、ギリシア文字や数学記号で表現しなさい。



3　対応のない二群の差の検定と単回帰分析

第 2 講で扱った対応のない t 検定を用いた独立した二群の差の検定と、この第 3 講で扱った単回帰モデルを比較した次の表とそれを解説した後続の文章を読み、空欄 A～H に入るのに適切な語句や文章を選択、あるいは、補いなさい。

	独立変数		RQ1 Yes/No	RQ2 How much
	数	タイプ		
二群の差の検定	A	C	E	G
単回帰分析	B	D	F	H

独立した二群の差の検定では、独立変数は[A　1 個だけであり・数に制限がなく]、単回帰分析では独立変数は[B　1 個だけである・数に制限がない]。また、二群の差の検定の独立変数は、[C　名義尺度・順序尺度・間隔尺度・比率尺度]に限られるが、単回帰

分析では、[D　名義尺度・順序尺度・間隔尺度・比率尺度]以下の全ての尺度データを扱うことができる。

どちらの枠組みであれ、大きく二つのリサーチクエスチョンがある。一つ目は、母集団におけるパラメータがゼロか否かを問うYes/no 疑問文であり、二群の差の検定では [　　E　　]というリサーチクエスチョンを考える。回帰分析で、これに対応するリサーチクエスチョンは[　　F　　]というものである。

二つ目は、母集団におけるパラメータがどのくらいなのかというHow much を問う疑問文であり、二群の差の研究では標準化効果量である[G $\delta =$　　]を考察するのに対して、回帰分析では非標準化効果量である[　　H　　]の値を検討対象に据える。

基本問題

4　重要な概念の確認

単回帰分析の統計的推測に関する重要概念を説明した次の各文章の記述が正しいかどうか、それぞれ正誤を述べなさい。

(1) $\varepsilon_i \sim N(0, \sigma^2)$のとき、$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$は直線を表す$(\sigma^2 \neq 0)$。

(2) xに「ドイツ語の履修経験があるか否か」という変数を採用する際には、ダミー変数によるコーディングが行われる。

(3) 回帰分析の係数の推定に最小二乗法が用いられるのは、この推定手法で得られる推定量に正規性と呼ばれる好ましい性質があるためである。

(4) xとyの相関係数が一定である場合、xの分散がyの分散より大きくなればなるほど、最小二乗法で推定された傾きを表す回帰係数$\hat{\beta}_1$の大きさ（絶対値）は大きくなる。

(5) 母集団における任意のパラメータθに対しその推定量を$\hat{\theta}$とする。このときこの推定量の期待値のことを$E[\hat{\theta}]$と表すとする。どんな推定量でも、$E[\hat{\theta}] = \theta$となることが知られている。

(6) 回帰係数の検定とは、母集団における直線がx軸に平行なのかどうかについて、*yes* か *no* かの判断を付けることである。

(7) データで得られたそれぞれのx_iの値に対して、それに対応する残差e_iがどのくらいになっているのかを図示したものが正規ＱＱプロットである。

(8) 予測値と残差の相関は必ず 1 になる。

(9) **決定係数**とは、独立変数がどれだけ従属変数の値を決定するかを示す指標である。

(10) **てこ値**が大きいものは、回帰係数の推定に強い影響を持ってしまっている可能性があるので、研究者がデータを眺めるなどしてしっかり検討・吟味することが必要だ。

5 統計モデル

問1 次の説明が正しければ○と、誤りであれば、×と記し、正しく訂正しなさい。

(1) $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$は母集団に想定される直線構造を表している。

(2)単回帰モデルで母集団に想定されているi番目のxの値とyの値が従う関係とは、直線と分散が等質の正規分布からなり、下記の書き方はこれを表している。

$y_i \sim N(\mu_i, \sigma_i^2),\ \mu_i = \beta_0 + \beta_1 x_i,\ \sigma_i^2 \sim N(0, 1)$

問2 単回帰モデルで母集団に想定されているi番目のxの値とyの値が従う関係は複数のやり方で表すことができる。考えられるだけ異なる書き方で表現しなさい。

6 統計モデル

次の文章を読み、後の問いに答えなさい。

研究者は、母集団に何らかのモデルを仮定して推論を行う。例えば、文の容認度に興味のある言語学者は「文に用いられている単語の数が長くなれば長くなるほど容認度判断が落ちるのではないか」という仮説を立てたとする。この仮説を数式で表したものが統計モデルである。

ただし、この仮説のままでは、まだ数式にはできない。なぜならば、単語の数の増加で文の容認度が落ちると言っても、それが直線的に落ちるのか、緩やかなカーブを描いて落ちるのか、いろいろなバリエーションがあるからだ。単回帰分析とは (a)二つの変数の間の関係に対して直線的な関係を想定する統計モデルであり、あまた想定されうる関係の中で、はなから (b)カーブを描くような関係を排除しているという点で、この統計モデルは母集団の関係に対して (c)かなり踏み込んだ強い仮定を持ち込んでいるとも言えるわけである。

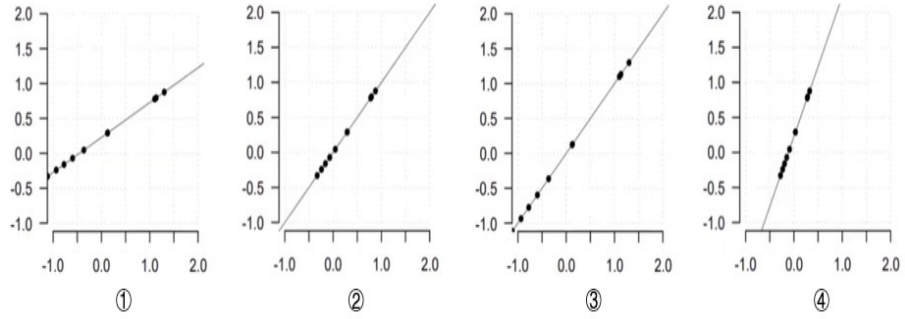
問１ 下線部(a)にあるように、母集団における直線を数式を使って表現しなさい。

問２ †下線部(b)にあるように、単回帰分析以外のモデルを母集団に想定することも可能である。母集団における何らかのカーブを想定している数式を一つ提案しなさい

問３ 下線部(c)にあるように、仮に真の母集団の分布が直線関係に基づいていないにもかかわらず単回帰分析を行った場合、残差に何らかの不整合性が見て取れる可能性がある。どのような不整合性が見られるか、具体的に説明しなさい。

7　回帰係数と相関係数

相関係数$\boldsymbol{r_{xy}}$を 1 とする場合、以下の図におけるそれぞれの直線の傾きの大きさについて、次の（A）から（D）の比較式のうち、正しいものを一つ選びなさい。



A. ①＜②＜③＜④　　B. ①＜②＝③＜④

C. ①＞②＞③＞④　　D. ①＞②＝③＞④

8　回帰係数と相関係数

あるドトールの店舗の、夏の 1 日の最高気温 $\boldsymbol{x_i}$(℃)とヨーグルンの売り上げ本数 $\boldsymbol{y_i}$(本)の関係についてのデータ $(\boldsymbol{x_i},\ \boldsymbol{y_i})$ $(1 \leqq \boldsymbol{i} \leqq 10)$を 10 個取ったら下表のようになった。

$\boldsymbol{x_i}$(℃)	$\boldsymbol{y_i}$(本)
26	105
28	112
28	113
29	112
29	117
31	118
31	123
32	117
33	123
33	130

この表におけるデータを使って計算すると、次のような結果になっている。

$\bar{x}$ ＝30,　$\bar{y}$＝ 117,　s_x^2＝5,　s_y^2＝45.2,　s_{xy}＝13.5

問１ 横軸を気温 **x**、縦軸を売り上げ本数**y**に対応させた散布図書き、各データをプロットしなさい。

問２ **x** と**y** の相関係数を計算しなさい。

問３ **y** の**x** に対する単回帰直線を求めなさい。そして、問１の図に回帰直線を描きなさい。

問４ 明日の最高気温が 30 ℃、明後日の最高気温が 27 ℃と予報があり、その通りになるとする場合、それぞれ何本分のヨーグルンの材料を準備しておくのか、単回帰分析で答えなさい。

（https://hiraocafe.com/note/simple-linear-regression.html によって改作）

９　回帰係数の算出

次の文章を読み、後の問いに答えなさい。

母集団に対して仮定された直線を推測するということは、その直線の傾きと切片に対して「合理的な値」を算出するということである。このように一つの値を計算して母集団のパラメータの値を求めることを[　A　]と呼ぶ。

標本に基づいて何らかの値を計算するということであるから、直線の傾きと切片に対する[　A　]を行うとは、すなわち、[　B　]を構成するということである。そして、[　A　]の目的で計算される[　B　]のことを[　C　]と呼ぶ。

しかし、「合理的な値」と理想を口にするのは簡単だが、それではいったい何を以って合理的と見なすのか、という点には当然厚みのある議論や緻密な検証が必要である。幸いなことに、卓越した統計学者たちの力によって、すでに多くのことが解明されている。例えば、(a)不偏性や不変性、(b)有効性などが[　C　]の持つ理想的性質として提案され、さらに、どういう[　C　]にこれらの特徴が見られるのか、見られないのかが研究、解明されてきた。

この[　C　]の選定の話を選挙にたとえてみよう。まず、研究者は候補者を擁立する。例えば、「これこれという方法で作り出した[　C　]を今回母集団のパラメータの値に使うことを提案したいのです」というようにである。しかし、別の研究者は別の[　C　]を提案するかもしれない。ちょうど（年齢さえ満たしていれば）誰でも選挙に立候補できるように、提案するだけならどのような[　C　]でも候補にできるのである。しかし、被選挙人

が立候補をしたからすぐさま採用されるわけではないのと同様に、提案された[　C　]が無批判的に採用されるわけではない。有権者の投票というハードルをクリアすることが候補者のふるい分けの役割を果たすのと同様に、この[　C　]の選択では不変性や不変性といった性質を満たすかどうかという点がフィルタリングの機能を果たすことになるのだ。

単回帰分析のような基礎的な統計モデルであれば、議論が尽くされているので「定石」とでもいうべき候補者が決まっている。(c)最小二乗法で計算した[　C　]がそれである。この計算結果は、実は別の方法、例えば最尤推定法に基づく[　C　]とも一致することが知られている。別々の方法でも同じ結果になるのであれば、より一層、最小二乗法の計算結果を母集団のパラメータの推定に使うことの妥当性としてみなすことができるだろう。

問１ 空欄 A～C に適切な用語を入れなさい。
問２ 下線部(a)に関して**不偏性**がどのようなものか説明しなさい。
問３ 下線部(a)に関して**不偏性**を持たない[　B　]にどのようなものがあるか例を挙げなさい。
問４ 下線部(b)に関して**有効性**がどのようなものか説明しなさい。
問５ 下線部(c)について**最小二乗法**がどのようなパラメータの推定方法なのかを説明しなさい。

10　決定係数

次の文章を読み、後の問いに答えなさい。

分散には大変便利な性質が知られており、ある変数yが互いに無相関な変数AとBの二つの和で表されるのであれば（すなわち$y = A + B$と書き表せるのであれば）、[　A　]という性質が成り立つことが知られている。これを[　B　]と呼ぶ。

回帰分析では、$y = \hat{y} + e$という関係が成り立つので、この性質を利用すると、[　C　]という関係が証明される。すなわち、データのばらつきが、予測値のばらつきと残差のばらつきに分解できるのである。

この性質を利用して、回帰直線がどれくらいデータにフィットしているのかを表す指標を組み立てることができる。(a)これが決定係数である。

ところで、なぜ[　B　]と呼ぶのであろうか。実は、これはs_yとs_Aとs_Bの間に[　D　]の定理が成り立つことに由来している。つまり、(b)これまで代数的な理解をしてきた数式に、私たちは幾何的な解釈を与えることができるのである。

問１ 空欄 A に当てはまる数式を次の中から選び記号で答えなさい。

① $s_y = s_A + s_B$　　② $s_y^2 = s_A^2 + s_B^2$

③ $s_y = s_A + 2s_A s_B + s_B$　　④ $s_y = s_A + 2s_A s_B + s_B$

⑤ $s_y^2 = s_A^2 + 2s_{AB} + s_B^2$　　⑥ $s_y^2 = s_A^2 + 2s_A s_B + s_B^2$

問２ 空欄 B から D を埋めなさい。

問３ 下線部(a)に関して、決定係数の定義を述べ、どのように空欄[　C　]の関係から決定係数が導かれるか説明しなさい。

問４ この[　　D　　]の定理をもとに、s_yとs_Aとs_Bを幾何的に表現しなさい。

実践問題

11　回帰分析の運用Ⅱ

あるデータに対して、次のような単回帰モデルを立ててその係数を最小二乗法によって推定した。その結果は下の表にまとめられている。これをもとに以下の問いに答えなさい。

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

	点推定値	標準誤差	t 値	p 値
β_0	60.30	4.25	[A]	< 2e-16
β_1	0.19	0.08	[B]	0.0149

残差の標準誤差: 11.82

サンプルサイズ: 47

決定係数: 0.02

問１ 空欄 A と B に入る数字を求めなさい。

問２ 有意水準を 0.05 とするとき、それぞれの係数の大きさがゼロであるか否かに対して統計的仮説検定を行うとき、どのような結論を出すことになるか、説明しなさい。

問３ 次の図１と図２はこの推定結果に基づいて算出した回帰直線の 95%信頼区間とデータの 95%予測区間のいずれかを表している。信頼区間を表しているのはどちらか答えなさい。

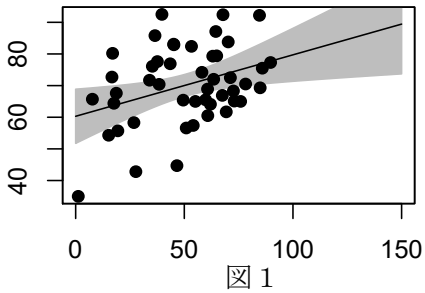


図１

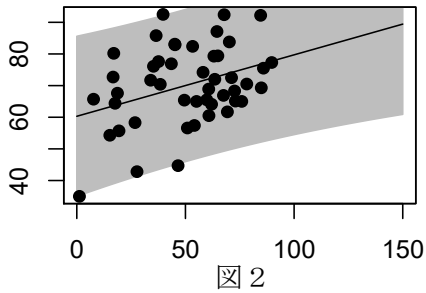


図２

問４ もし未来のデータを得た時に、その x の値が 50 であり、かつ y の値が 50 であるデータは起こりやすいものと考えることができるのか、適切に応答せよ。

問５ 今回作成した回帰モデルはデータにフィットしていると言えると言えるかどうか、適切に応答せよ。

問６ 単回帰分析を行う際には、母集団についていくつかの仮定を設けている。しかし、それが満たされているか否かを検討する必要は当然存在する。そこで、下に回帰診断に用いられる様々なプロットを描画した。ここからどのようなことが読み取れ、誠実な研究者としてどのような検討を行う必要があるか、すでに提示されている図 1、図２と併せながら分かりやすく説明しなさい。

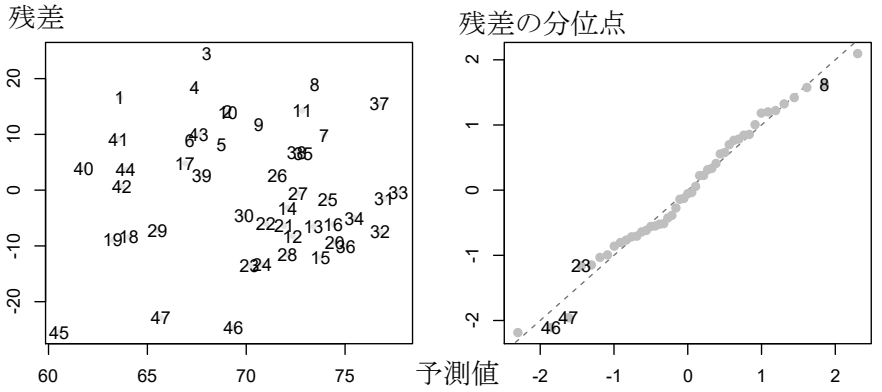


図３残差プロット

図４　正規 QQ プロット

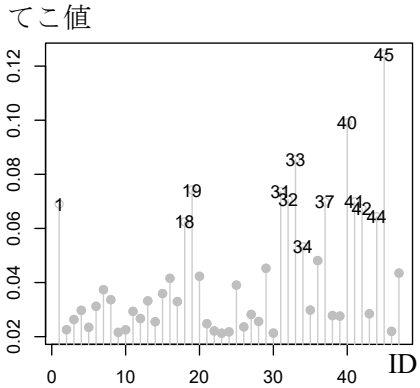


図５　てこ値

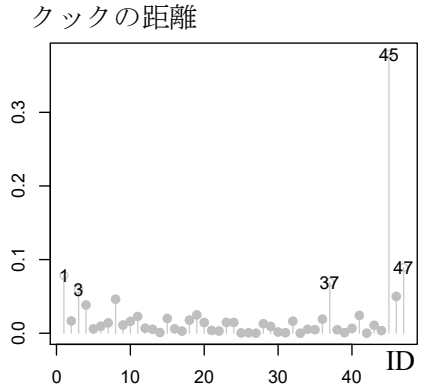


図６クックの距離

※データにおける数字は、データの値につけられた通し番号（ID）を表している（*i* 番目のデータと呼ぶときの *i* の値である。なお、一部の図では作図の都合上、小さい値は●で表し数字を省略している。

12　回帰分析の運用Ⅱ

あるデータに対して、次のような単回帰モデルを立ててその係数を最小二乗法によって推定した。その結果は下の表にまとめられている。これをもとに以下の問いに答えなさい。

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

	点推定値	標準誤差	t 値	p 値
β_0	1.34	0.65	[A]	0.07
β_1	0.70	0.10	[B]	0.000144

残差の標準誤差: 0.94
サンプルサイズ: 10
決定係数: 0.72

問１ 空欄 A と B に入る数字を求めなさい。

問２ 有意水準を 0.05 とするとき、それぞれの係数の大きさがゼロであるか否かに対して統計的仮説検定を行うとき、どのような結論を出すことになるか、説明しなさい。

問３ 今回作成した回帰モデルはデータにフィットしていると言えると言えるかどうか、適切に応答せよ。

問４ 単回帰分析を行う際には、母集団についていくつかの仮定を設けている。しかし、それが満たされているか否かを検討する必要は当然存在する。そこで、下に信頼区間と予測区間（点線と灰色のエリアのどちらがどちらかは適切に判断すること）、および、回帰診断に用いられる様々なプロットを描画した。ここからどのようなことが読み取れ、誠実な研究者としてどのような検討を行う必要があるか、分かりやすく説明しなさい

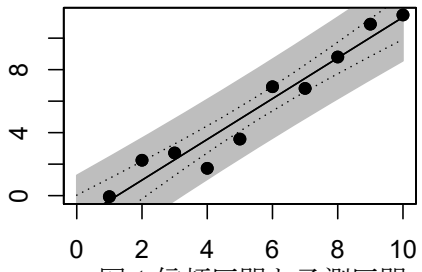


図１信頼区間と予測区間
（●はデータの位置を表す）

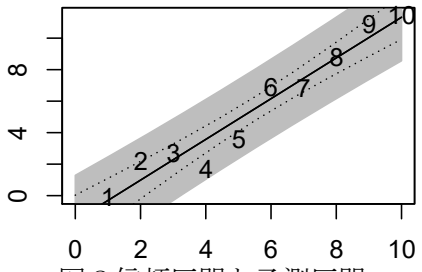


図２信頼区間と予測区間
（数字はデータの ID を表す）

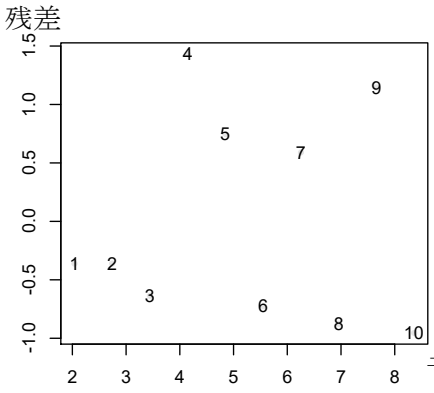


図３残差プロット

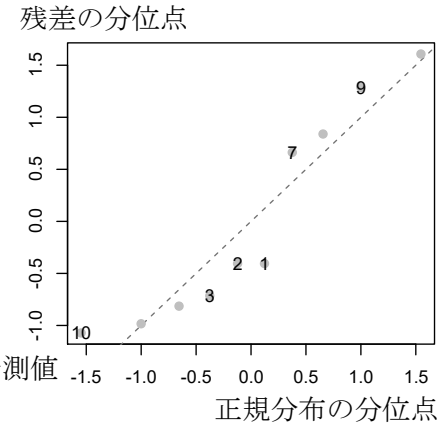


図４　正規 QQ プロット

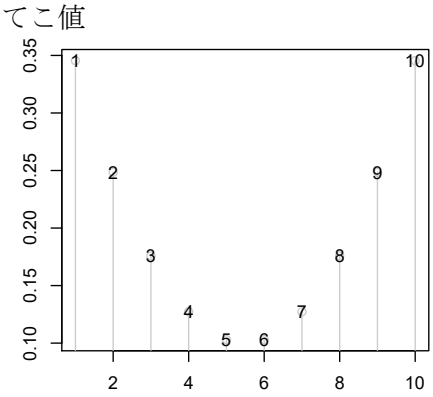


図５　てこ値

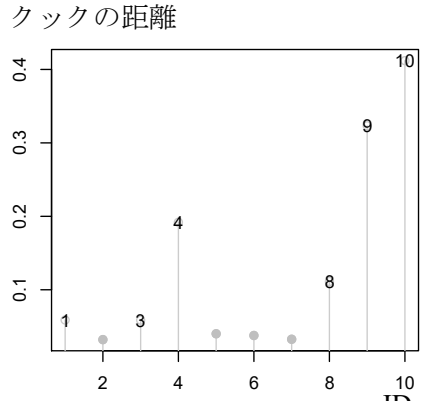


図６クックの距離

※データにおける数字は、データの値につけられた通し番号（ID）を表している（*i* 番目のデータと呼ぶときの *i* の値である。なお、一部の図では作図の都合上、小さい値は●で表し数字を省略している。

13　独立変数と従属変数

日本の小学生を母集団として、全国の小学生からランダムに人を選んで、独立変数を体重（kg）、従属変数を年齢（歳）に据えて、回帰分析を行った。サンプルサイズは 10 であり、傾きを表す回帰係数を見ると、それがゼロであるという帰無仮説は有意に棄却でき、その点推定値は 2.25 であった。回帰診断でもなんら問題点は見つからなかったとする。

さて、この統計解析をもとに、ある研究者が次のように結論を下したとする。

点推定値が 2.25 ということは、体重 1kg 上がると年齢が 2 年 3 か月上昇するということであり、日本の学校制度では体重によって学年が決まるのである。

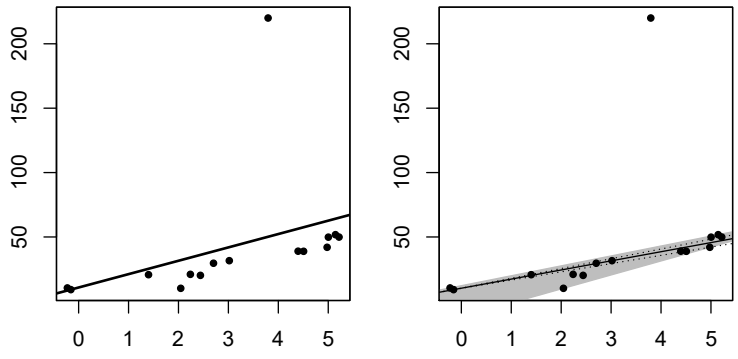
もちろん、研究者の結論は正しくないわけであるが、(単) 回帰分析という統計手法を説明しながら、この推論のどこが間違っているのかを的確に指摘せよ。

14 †分位点回帰

単回帰分析のような基礎的な統計モデルであれば、そのパラメータの推測方法には「定石」とでもいうべき候補者が決まっている。どんな教科書でも最小二乗法が紹介され、その計算結果が最尤推定法を用いた結果と一致することは、よく知られている。

しかしながら、最小二乗法には弱点も知られている。最小二乗法とは、[　A　]という統計量の計算にも用いられている指標だが、[　A　]が[　B　]などの統計指標と比べて[　C　]に弱いことは、この統計指標の弱点として考えられることが多い。

同じように、最小二乗法によって計算された回帰直線も[　C　]の影響を強く受けてしまうという弱点がある。[　C　]がある場合、それに引きずられて、データが集まっている場所から離れたところに推定された直線が引かれてしまうことがあるのである。例えば、下記の図のうち左のものはデータに最小二乗法で直線をあてはめたものであり、直線がデータから外れていることが読み取れるだろう。



そこで、ちょうど[　B　]が[　A　]という統計量とは異なり、[　C　]に強いように、最小二乗法を使わずに回帰直線を推定するという発想が生まれてくる。[　C　]と思われるデータが存在する場合に、最小二乗法に代わってしばしば用いられるの

が分位点回帰（ロバスト回帰）であり、x の値が与えられたときの y の条件付分位点をモデル化したものである。上記に示された右図がこれを表し、灰色の部分が 90%分位点、点線が 66%分位点、真ん中の実線が 50%分位点（すなわち[　B　]の構造）をトレースしている。[　C　]の影響を受けず、左図のときよりもデータの真ん中を捉えていることが分かるだろう。

このように、最小二乗法には[　C　]に弱いという特徴があるので、最小二乗法を用いた通常の単回帰分析を行う際には、回帰係数の推定に大きな影響を与えるような[　C　]の存在に敏感になることが求められる。上記図に載せたような x 軸の値としては他の値と変わらない一方で、y 軸方向には極端に大きな値を取るものは、例えば[　D　]と呼ばれる指標で探知することができる。特に明確な閾値があるわけではないが、例えば、この値の値が 0.5 を超えるようであれば、最小二乗法を用いた推定結果には慎重な態度を取ることが望ましいであろう。分位点回帰を併用したり、場合によっては[　C　]を例外扱いをして、それ以外のデータのみに回帰分析をあてがうなど、柔軟にデータにむきあうことが大切である。