統計学サブノート

上野山　清久

私の統計学の学習成果を少しずつまとめていこうと思います。

入門書の数式を丹念に追うことにより、最近少しずつですが統計学に対する理解が進んできたと感じています。

ただし理解できたと思っても、しばらく経ってから読み返してみるとその内容をすっかり忘れてしまっていたり、解った気分になっていただけで実はよく解っていなかったということに気づいたりして、同じところを丹念に読み直さなければならないということも多々あります。

やはり物事を深く理解するためには、自分なりにその内容を整理してみること、そしてその際に他の誰かに説明することを意識しながらまとめてみることが効果的なのではないかと思います。以前何かのテレビ番組でジャーナリストの池上彰さんが、“本を読む際はその内容を自分が他人に説明することを意識しながら読むように心がけている”というお話をされているのをお聞きしたことがあります。それ以来私もこれを実践するようにしていますが、そうした意識で本を読むようになってから以前よりその内容がより理解できるようになったと感じています。

また最近勉強していて気づいたことですが、本の内容が理解できないのには主に２つの要因があるように思います。それは、（１）そもそもその内容が自分の理解レベルを超えている場合と、（２）現在の自分の実力で十分理解できる内容であるにもかかわらず、何かしらの思い込み、思い違い等がある場合です（もちろん、その文章の内容が間違っているという場合もあるでしょうが、ここではそれは除外します。）。

（１）の場合は、そのレベルに追いつくために日々努力する以外に方法はありませんが、案外（２）が原因の場合も多いのではないでしょうか。この場合は、自分の思考パターンの癖を意識しながら丹念に文章を読み、自分が引っ掛かっている箇所に気づく以外に方法はないと思います。

（１）の場合は、そこに書かれている内容についていけないことがすぐに分かるでしょうから、何となく解ったようで解らないという場合は（２）が原因のことが多いのではないかと思います。

このようなことも意識しながら統計学を勉強していく過程で私が理解した内容を整理しておいた方がよいと感じたことを思いつくままに書いていく予定なので内容に一貫性があるわけではなく、そういう意味ではサブノートというより雑記帳という方がふさわしいのかもしれません。

ホームページで公開する以上は第三者の目に触れることも意識しながら作成するつもりですが、とりあえずは私的なサブノートに過ぎませんので、私以外の人が読んでも役に立たないかもしれませんが、随時加筆・修正していくことにより徐々に内容を充実させていき、ゆくゆくは多くの人に理解しやすいと感じていただける統計学サブノートにしていきたいと思っています。

2017年　6月 4日

（最新更新日　2017年　7月 1日）

○分散はなぜ２乗するのか

n 個のデータ $X_i\left(i=1,2,\cdots,n\right)$ の平均値を $\overline{X}$ とすると、

X_i の標本分散：$S^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$　・・・①

任意の定数を a とすると、偏差平方和 $\sum_{i=1}^{n}\left(X_i-a\right)^2$ は、$a=\overline{X}$ のときに最小値をとる。

すなわち、$\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 \le \sum_{i=1}^{n}\left(X_i-a\right)^2$ (※)　・・・②

両辺に $\frac{1}{n}$ を掛けると、$S^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 \le \frac{1}{n}\sum_{i=1}^{n}\left(X_i-a\right)^2$　・・・②’ が成り立つ。

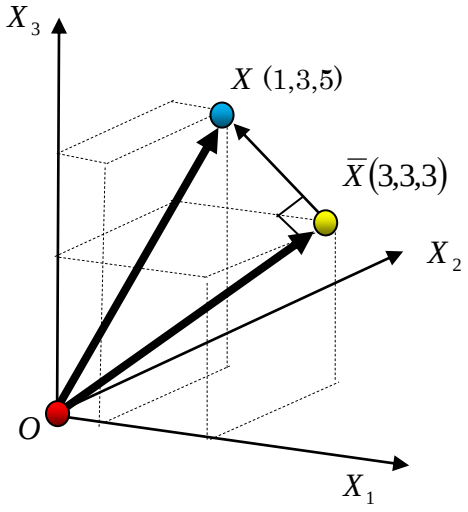
話を単純化し、３個の具体的なデータ $X=\left(1,3,5\right)$ で考えると、平均値 $\overline{X}=3$、偏差平方和は $\left(1-3\right)^2+\left(3-3\right)^2+\left(5-3\right)^2=8$ である。ちなみに例えば、$a=2$ とするとその偏差平方和は 11 となり確かに②式および②’式を満たしている。

X および $\overline{X}$ を３次元ベクトル空間における位置ベクトルとして表示すると、$\overrightarrow{OX}=\left(1,3,5\right),\overrightarrow{O\overline{X}}=\left(3,3,3\right)$ であり、これらの偏差ベクトル $\overrightarrow{\overline{X}X}=\left(-2,0,2\right)$ である〔図１〕。

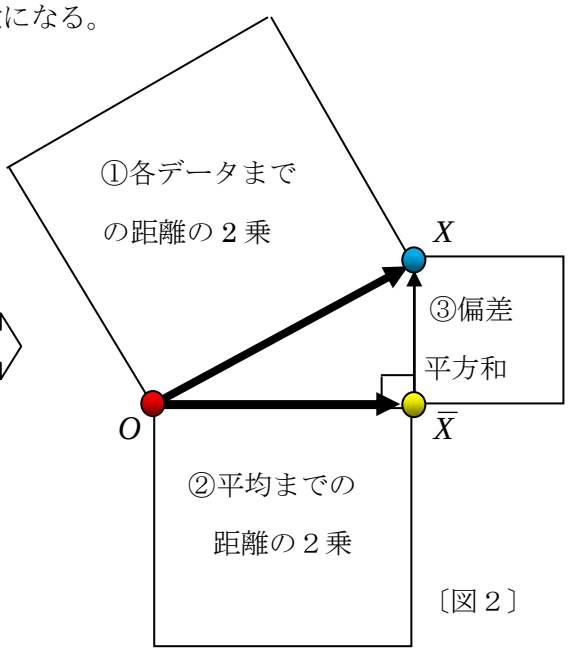
このとき、△$O\overline{X}X$ は、線分 OX を斜辺とする直角三角形である。

したがって、三平方の定理より $\left|\overline{X}X\right|^2=\left|OX\right|^2-\left|O\overline{X}\right|^2$ は成り立つが、$\left|\overline{X}X\right|=\left|OX\right|-\left|O\overline{X}\right|$ は成り立たないため、偏差で考えるより偏差の２乗で考える方が好都合である〔図２〕。

そして、偏差平方和をデータ数で割ると分散になる。



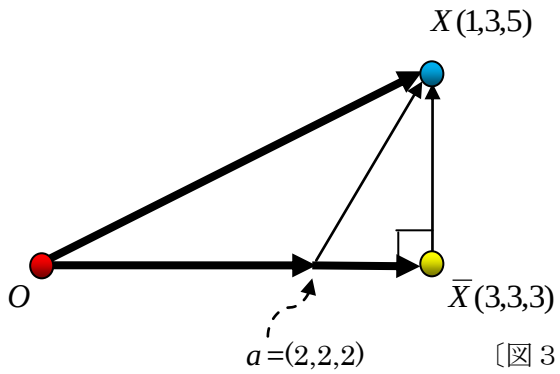
〔図１〕



〔図２〕

なお、データが４個以上の場合は、もはや三次元空間では図示できないが、考え方は同じである。

△$O\bar{X}X$ が直角三角形になるという点について補足すると、$\bar{X}$=3（位置ベクトル表示では(3,3,3)）のとき$\overrightarrow{O\bar{X}}$ と $\overrightarrow{\bar{X}X}$ が直交するため偏差が最小となり、例えば$a=2$（同(2,2,2)）のときの偏差はこれより大きくなる〔図３〕。



〔図３〕

〈凡人のメモ〉
最小２乗法はなぜ２乗するのかについてもこれと全く同じ考え方で、〔図２〕の①「各データまでの距離の２乗」を「各データから平均値までの距離の２乗」に、②「平均までの距離の２乗」を「理論値から平均値までの距離の２乗」に、③「偏差平方和」を「残差平方和」にそれぞれ読みかえれば良いのですが、〔図１〕と同じように図示するとなると複雑な図になることに気づきました（マイナスの座標も必要になるため）。これについては、別の機会に説明しようと思います。

〈凡人のメモ〉
各データとの偏差の平方和を最小にする値が平均値であるということは感覚的に受け入れやすいのですが、各データとの偏差の絶対値の和を最小にする値は、実は平均値ではなく中央値であると事実はどうも感覚的にしっくり来ないと個人的に感じています。このことが私が分散などの概念を深く理解するのを困難にしている一因なのかもしれません。

(※) $$\sum_{i=1}^{n}(X_i-a)^2=\sum_{i=1}^{n}(X_i-\bar{X}+\bar{X}-a)^2$$

$$=\sum_{i=1}^{n}(X_i-\bar{X})^2+\sum_{i=1}^{n}(\bar{X}-a)^2+\sum_{i=1}^{n}2(X_i-\bar{X})(\bar{X}-a)$$

$$=\sum_{i=1}^{n}(X_i-\bar{X})^2+n(X_i-a)^2+2(\bar{X}-a)\underbrace{\sum_{i=1}^{n}(X_i-\bar{X})}_{0}$$

$$=\sum_{i=1}^{n}(X_i-\bar{X})^2+\underbrace{n(X_i-a)^2}_{\geq 0}\geq\sum_{i=1}^{n}(X_i-\bar{X})^2$$

○不偏分散はなぜ$n-1$で割るのか

統計学の入門書でしばしば不偏分散はnではなく$n-1$で割る理由について、「自由度が１少ないから」という説明をしていますが、この説明が統計学を一般人から遠ざけている一因ではないのだろうかと最近思います。「自由度が１少ないからそうなのだと言われても納得できないが、これで納得できない自分には統計学の理解は無理なんだろう。」と思ってしまう人も少なからずいるのではないでしょうか。

１　期待値と分散

（１）確率変数の平均値（期待値）

期待値とは、将来実現する可能性のある数値の平均のことである。

確率変数Xがn通りの値をとる可能性があり、その実現値をX_iとし、X_iが実現する確率をP_iとすると、Xの期待値$E[X]$は、

$$E[X]=\sum_{i=1}^{n} X_i P_i = X_1P_1+X_2P_2+\cdots+X_nP_n$$

例えば、サイコロは１から６までの６通りの目が出る可能性があり、それぞれの目が出る確率は各６分の１と考えられる。

したがって、サイコロを振る際の期待値は、

$$E[X]=\sum_{i=1}^{6} X_i P_i = X_1P_1+X_2P_2+X_3P_3+X_4P_4+X_5P_5+X_6P_6$$

$$=1\times\frac{1}{6}+2\times\frac{1}{6}+3\times\frac{1}{6}+4\times\frac{1}{6}+5\times\frac{1}{6}+6\times\frac{1}{6}=3.5$$

実際にサイコロを振った結果はこれと異なることも多く、例えば６回振ってその実現値がたまたますべて１だったとしたらその平均値は１ということになり、期待値の３．５とは異なることになる。

・期待値に関する重要な性質

(1)　$E[c]=c$（ただしcは定数）

(2)　$E[X+c]=E[X]+c$

(3)　$E[cX]=cE[X]$

(4)　$E[X\pm Y]=E[X]\pm E[Y]$

〈凡人のメモ〉

よく「期待値＝平均」であるという説明がされますが、それなら、なぜわざわざ「平均」と区別して「期待値」と言うのか。私はそんなことが疑問で期待値というものが今一つ理解できていませんでしたが、ある時、要するに平均のうち、まだ結果が出ていない（未来に起こりうる）数値の平均のことを期待値というのだということに気づいてやっと納得した覚えがあります。そんなことは参考書を見れば書いてあることなんですが、自分のアンテナにはなかなか引っ掛かってくれないものです。

（２）確率変数の分散

確率変数の分散とは、データの期待値からの散らばり具合を示す数値であり、期待値からの偏差の２乗和の期待値のことである。確率変数 X の分散を $V[X]$ とし、

$E[X]=\mu$ とすると、$V[X]=E\left[(X-\mu)^2\right]$

・分散に関する重要な性質

(1) $V[c]=0$

(2) $V[X+c]=V[X]$

(3) $V[cX]=c^2V[X]$

(4) $V[X\pm Y]=V[X]+V[Y]$

〈凡人のメモ〉

確率変数の平均値のことを「期待値」と呼んで区別するなら、確率変数の分散も名称を区別すべきだと思いますが、なぜか分散は確率変数であろうがなかろうがどちらも分散と呼んで区別しないのはなぜなのでしょう。あえて区別するとすれば、確率変数の分散は「期待値からの偏差の２乗和の期待値」ということになりますが、これではちょっと長過ぎるので、「期待分散」とでも呼べばよいのではないでしょうか。

以上を整理すると、実現値の平均値＝平均値、確率変数の平均値＝期待値

実現値の分散　＝分散、　確率変数の分散　＝期待値からの偏差の２乗和の期待値

２　$n-1$ で割る理由

n 個のデータ $X_i\left(i=1,2,\cdots,n\right)$ の平均値を $\overline{X}$ とすると、X_i と任意の定数 a との偏差平方和 $\sum_{i=1}^{n}\left(X_i-a\right)^2$ は、$a=\overline{X}$ のときに最小値をとる。

したがって、母平均を μ とすると、

$\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\le\sum_{i=1}^{n}\left(X_i-\mu\right)^2$ が成り立つ。

ゆえに両辺を n で割って、

$\underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}_{\text{標本分散 } S^2}\le\underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu\right)^2}_{\text{母分散 } \sigma^2}$　となるため、標本平均 $\overline{X}$ が母平均 μ と偶然一致

する場合を除いて標本分散S^2は母分散σ^2より小さい値となる。

それでは、S^2をどれだけ水増しすればσ^2の不偏分散s^2になるのか。

$$S^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu+\mu-\overline{X}\right)^2=\frac{1}{n}\sum_{i=1}^{n}\left\{\left(X_i-\mu\right)-\left(\overline{X}-\mu\right)\right\}$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu\right)^2+\frac{1}{n}\underbrace{\sum_{i=1}^{n}\left(\overline{X}-\mu\right)^2}_{n\left(\overline{X}-\mu\right)^2}-2\frac{1}{n}\sum_{i=1}^{n}\left\{\left(X_i-\mu\right)\left(\overline{X}-\mu\right)\right\}$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu\right)^2+\frac{1}{n}\cdot n\left(\overline{X}-\mu\right)^2-2\left(\overline{X}-\mu\right)\frac{1}{n}\underbrace{\sum_{i=1}^{n}\left(X_i-\mu\right)}_{n\left(\overline{X}-\mu\right)}$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu\right)^2+\left(\overline{X}-\mu\right)^2-2\left(\overline{X}-\mu\right)^2$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu\right)^2-\left(\overline{X}-\mu\right)^2$$

$$E\left[S^2\right]=E\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right]$$

$$=E\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu\right)^2-\left(\overline{X}-\mu\right)^2\right]$$

$$=E\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\mu\right)^2\right]-E\left[\left(\overline{X}-\mu\right)^2\right]$$

$$=\frac{1}{n}E\left[\left(X_1-\mu\right)^2+\left(X_2-\mu\right)^2+\cdots+\left(X_n-\mu\right)^2\right]-E\left[\left(\overline{X}-\mu\right)^2\right]$$

$$=\frac{1}{n}\left\{E\left[\left(X_1-\mu\right)^2\right]+E\left[\left(X_2-\mu\right)^2\right]+E\left[\left(X_n-\mu\right)^2\right]\right\}-E\left[\left(\overline{X}-\mu\right)^2\right]$$

$$=\frac{1}{n}\left(\underbrace{V\left[X_1\right]+V\left[X_2\right]+\cdots+V\left[X_n\right]}_{n\sigma^2(※1)}\right)-\underbrace{V\left[\overline{X}\right]}_{\frac{\sigma^2}{n}(※2)}$$

$$=\frac{1}{n}n\sigma^2-\frac{\sigma^2}{n}$$

$$=\frac{n-1}{n}\sigma^2 \quad \cdots \quad E\left[S^2\right] \text{は} \sigma^2 \text{より小さい。}$$

両辺に$\frac{n}{n-1}$を掛けると、

$$E\left[S^2\right]\times\frac{n}{n-1}=E\left[\frac{1}{n}\sum\left(X_n-\bar{X}\right)^2\right]\times\frac{n}{n-1}=E\left[\frac{1}{n-1}\sum\left(X_n-\bar{X}\right)^2\right]=E\left[s^2\right]=\sigma^2$$

$$\therefore s^2=\frac{1}{n-1}\sum\left(X_n-\bar{X}\right)^2$$

(※1) $\frac{1}{n}\left(V\left[X_1\right]+V\left[X_2\right]+\cdots+V\left[X_n\right]\right)$

$$=\frac{1}{n}\left(\sigma^2+\sigma^2+\cdots+\sigma^2\right)=\frac{1}{n}n\sigma^2=n\sigma^2$$

(※2) $V\left[\bar{X}\right]=V\left[\frac{X_1+X_2+\cdots+X_n}{n}\right]=\frac{1}{n^2}V\left[X_1+X_2+\cdots+X_n\right]$

$$=\frac{1}{n^2}\left(V\left[X_1\right]+V\left[X_2\right]+\cdots+V\left[X_n\right]\right)=\frac{1}{n^2}n\sigma^2=\frac{\sigma^2}{n}$$

○t分布について

統計学を学び始めて多くの人が最初に越え難い壁と感じるのがt分布ではないでしょうか。正規分布は同年齢の人の身長が従う分布であるなどの具体例で考えるとイメージしやすいのに対して、t分布はなかなかイメージしづらく、そもそも正規分布と何がどう違うのかが理解しにくいのではないかと思います。

母分散が既知の場合の母平均の推定についての説明からはじまってt分布へたどり着くまでの道のりが長いことや、「t分布」といいながら、なぜか標準正規分布やχ^2分布という別の分布がからんでくること、さらに自由度というこれまたよく解らない概念が出てきて、その上自由度nのケースの話をしているかと思うと、気がついたら自由度$n-1$の話になっていたりするなど、いろいろな概念が入り乱れている割には、結局は標準正規分布によく似た分布だという、説明が回りくどくて解りにくい上につかみどころのない分布だなぁというのが私の印象です。

1　χ^2（カイ２乗）分布

（１）標準正規分布$N(0,1)$に従う互いに独立な確率変数$Z_1, Z_2, \cdots, Z_n$の各項の２乗和、

$Z_1^2+Z_2^2+\cdots+Z_n^2$が従う確率分布を自由度nのχ^2分布といい、$\chi^2(n)$と表す。

$$\sum_{i=1}^{n} Z_i^2 \sim \chi^2(n) \quad \cdots ①$$

（２）（１）より、正規分布$N(\mu,\sigma^2)$に従う互いに独立な確率変数をそれぞれ標準化した

$\dfrac{X_1-\mu}{\sigma}, \dfrac{X_2-\mu}{\sigma}, \cdots, \dfrac{X_n-\mu}{\sigma}$の各項の２乗和、

$\left(\dfrac{X_1-\mu}{\sigma}\right)^2+\left(\dfrac{X_2-\mu}{\sigma}\right)^2+\cdots+\left(\dfrac{X_n-\mu}{\sigma}\right)^2$は、自由度$n$の$\chi^2$分布に従う。

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2(n) \quad \cdots ②$$

（３）不偏分散：$s^2=\dfrac{(X_1-\bar{X})^2+(X_2-\bar{X})^2+\cdots+(X_n-\bar{X})^2}{n-1}$の両辺にそれぞれ$n-1$を掛けて$\sigma^2$で割ると、

$$\frac{(n-1)s^2}{\sigma^2}=\left(\frac{X_1-\bar{X}}{\sigma}\right)^2+\left(\frac{X_2-\bar{X}}{\sigma}\right)^2+\cdots+\left(\frac{X_n-\bar{X}}{\sigma}\right)^2$$

となるが、これは自由度$n-1$のχ^2分布に従う。これをYとおくと、

$$Y=\frac{(n-1)s^2}{\sigma^2}=\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2 \sim \chi^2(n-1) \quad \cdots ③$$

※ ②のμを$\bar{X}$に入れ替えた式が③である。

$s^2=\dfrac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$　両辺にそれぞれ$n-1$を掛けて、

$$(n-1)s^2=\sum_{i=1}^{n}(X_i-\bar{X})^2$$

$$=\sum_{i=1}^{n}\{X_i-\mu-(\bar{X}-\mu)\}^2$$

$$=\sum_{i=1}^{n}\{(X_i-\mu)^2-2(X_i-\mu)(\bar{X}-\mu)+(\bar{X}-\mu)^2\}$$

$$=\sum_{i=1}^{n}(X_i-\mu)^2-2(\bar{X}-\mu)\underbrace{\sum_{i=1}^{n}(X_i-\mu)}_{n(\bar{X}-\mu)}+\underbrace{\sum_{i=1}^{n}(\bar{X}-\mu)^2}_{n(\bar{X}-\mu)^2}$$

$$=\sum_{i=1}^{n}(X_i-\mu)^2-2n(\bar{X}-\mu)^2+n(\bar{X}-\mu)^2$$

$$=\sum_{i=1}^{n}(X_i-\mu)^2-n(\bar{X}-\mu)^2$$

両辺をσ^2で割ると、

$$\frac{(n-1)s^2}{\sigma^2}=\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2-\underbrace{\frac{n(\bar{X}-\mu)^2}{\sigma^2}}_{\left(\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}\right)^2}$$

右辺第1項は$\chi^2(n)$、第2項は$\chi^2(1)$より、

$$\frac{(n-1)s^2}{\sigma^2}=\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2 \sim \chi^2(n-1) \quad \cdots ③と同じ$$

〈凡人のメモ〉

$\frac{(n-1)s^2}{\sigma^2}$ ではなく $\frac{ns^2}{\sigma^2}$ で説明している参考書もありますが、その場合も自由度は $n-1$ です。$\frac{(n-1)s^2}{\sigma^2}$ の分子の $n-1$ が n に代わると自由度も n になると思ってしまう人（実は私）もいると思いますが、そうではなく自由度は n 個の項のうち自由に動ける項の数で決まります。

2　t 分布

$$Z=\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$$　・・・④　は標準正規分布 $N(0,1)$ に従う。

しかし、④の σ^2 を s^2 で代用した

$$t=\frac{\bar{X}-\mu}{\sqrt{\frac{s^2}{n}}}$$　・・・⑤　は標準正規分布 $N(0,1)$ には従わず、自由度 $n-1$ の t 分布に従う。

t 分布がどのような分布かを調べるために、⑤を以下のように変形する。

$$t=\frac{\bar{X}-\mu}{\sqrt{\frac{s^2}{n}}}=\frac{\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}}{\frac{\sqrt{\frac{s^2}{n}}}{\sqrt{\frac{\sigma^2}{n}}}}=\frac{\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}}{\frac{\sqrt{s^2}}{\sigma^2}}=\frac{\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}}{\sqrt{\frac{(n-1)s^2}{\sigma^2}\cdot\frac{1}{n-1}}}=\frac{Z}{\sqrt{\frac{Y}{n-1}}}$$

分母・分子をそれぞれ $\sqrt{\frac{\sigma^2}{n}}$ で割る。

分母・分子にそれぞれ $\sqrt{n}$ をかける。

$\sqrt{\ }$ の中の分母・分子にそれぞれ $\frac{n-1}{\sigma^2}$ をかける。

要するに標準正規分布 (Z) を、自由度 $n-1$ の χ^2 分布 (Y) を自由度で割って平方根をとったもので割ると t 分布になる。

○回帰分析について

式展開を地道に追うことにより回帰分析に対する理解が深まりつつあると感じていますが、あまりにも数式が多いのでこれを読んでくれる人は少ないだろうとは思いますが・・・。

1　偏差平方和（偏差２乗和）と偏差積和

（1）偏差平方和

$$S_{xx}=\sum\left(X_i-\bar{X}\right)^2=\sum\left(X_i^2-2X_i\bar{X}+\bar{X}^2\right)$$
$$=\sum X_i^2-2\bar{X}\underbrace{\sum X_i}_{n\bar{X}}+\underbrace{\sum\bar{X}^2}_{n\bar{X}^2}$$
$$=\sum X_i^2-2n\bar{X}^2+n\bar{X}^2$$
$$=\sum X_i^2-n\bar{X}^2 \quad \cdots (1\text{-}1)$$

また、

$$S_{xx}=\sum\left(X_i-\bar{X}\right)^2=\sum\left(X_i-\bar{X}\right)\left(X_i-\bar{X}\right)$$
$$=\sum\left\{\left(X_i-\bar{X}\right)X_i-\left(X_i-\bar{X}\right)\bar{X}\right\}$$
$$=\sum\left(X_i-\bar{X}\right)X_i-\bar{X}\underbrace{\sum\left(X_i-\bar{X}\right)}_{0}$$
$$=\sum\left(X_i-\bar{X}\right)X_i \quad \cdots (1\text{-}2)$$

$$S_{xx}=\sum\left(X_i-\bar{X}\right)^2=\sum X_i^2-n\bar{X}^2 \quad \cdots (1\text{-}1)$$
$$=\sum\left(X_i-\bar{X}\right)X_i \quad \cdots (1\text{-}2)$$

（2）偏差積和

$$S_{xy}=\sum\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)=\sum\left(X_iY_i-X_i\bar{Y}-\bar{X}Y_i+\bar{X}\bar{Y}\right)$$
$$=\sum X_iY_i-\bar{Y}\sum X_i-\bar{X}\sum Y_i+\sum\bar{X}\bar{Y}$$
$$=\sum X_iY_i-n\bar{X}\bar{Y}+n\bar{X}\bar{Y}+n\bar{X}\bar{Y}$$
$$=\sum X_iY_i-n\bar{X}\bar{Y} \quad \cdots (1\text{-}3)$$

また、

$$S_{xy}=\sum\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)=\sum\left\{\left(X_i-\bar{X}\right)Y_i-\left(X_i-\bar{X}\right)\bar{Y}\right\}$$
$$=\sum\left(X_i-\bar{X}\right)Y_i-\bar{Y}\underbrace{\sum\left(X_i-\bar{X}\right)}_{0}$$
$$=\sum\left(X_i-\bar{X}\right)Y_i \quad \cdots (1\text{-}4)$$

$$S_{xy}=\sum(X_i-\bar{X})(Y_i-\bar{Y})=\sum X_iY_i-n\bar{X}\bar{Y} \quad \cdots (1\text{-}3)$$
$$=\sum(X_i-\bar{X})Y_i \quad \cdots (1\text{-}4)$$

2　最小２乗法

最小２乗法は、各データとの y 軸方向の距離（残差）の２乗を最小にするような直線を求める方法ですが、学び始めの頃はなぜ x 軸方向の距離ではだめなんだろうと思っていました。X（説明変数）の値によって Y（被説明変数）の値がどう変化するかを知りたいのだから当然なんですが、私と同じ疑問を持つ人も少なからずいるのではないでしょうか。

最小２乗法によって求められる回帰直線　$\hat{Y}_i=\hat{\alpha}+\hat{\beta}X_i$　の $\hat{Y}_i$（理論値）と、実際に観測された各データ Y_i（実績値）との差のことを残差（$\hat{u}_i$）という。

最小２乗法は、この残差の２乗和 $\sum\hat{u}_i^2$ を最小にするような直線の係数である $\hat{\alpha}$ と $\hat{\beta}$ を求める方法である。

残差：$\hat{u}_i=\underbrace{Y_i}_{(\text{実績値})}-\underbrace{\hat{Y}_i}_{(\text{理論値})}=Y_i-(\hat{\alpha}+\hat{\beta}X_i)$

残差２乗和：$\sum\hat{u}_i^2=\sum\{Y_i-(\hat{\alpha}+\hat{\beta}X_i)\}^2 \quad \cdots (2\text{-}1)$

(2-1)を $\hat{\alpha}$、$\hat{\beta}$ でそれぞれ偏微分すると、

$$\frac{\partial\sum\hat{u}_i^2}{\partial\hat{\alpha}}=-2\sum\underbrace{\{Y_i-(\hat{\alpha}+\hat{\beta}X_i)\}}_{\hat{u}_i}=0 \quad \cdots (2\text{-}2) \quad \therefore\sum\hat{u}_i=0$$

$$\frac{\partial\sum\hat{u}_i^2}{\partial\hat{\beta}}=-2\sum X_i\underbrace{\{Y_i-(\hat{\alpha}+\hat{\beta}X_i)\}}_{\hat{u}_i}=0 \quad \cdots (2\text{-}3) \quad \therefore\sum\hat{u}_iX_i=0$$

(2-2)、(2-3)をそれぞれ整理すると、

$$\left.\begin{aligned}&\sum Y_i=n\hat{\alpha}+\hat{\beta}\sum X_i \quad \cdots (2\text{-}2)'\\&\sum X_iY_i=\hat{\alpha}\sum X_i+\hat{\beta}\sum X_i^2 \quad \cdots (2\text{-}3)'\end{aligned}\right\}\text{ 正規方程式}$$

(2-2)' より

$$\hat{\alpha}=\frac{\overbrace{\sum Y_i}^{n\bar{Y}}}{n}-\frac{\hat{\beta}\overbrace{\sum X_i}^{n\bar{X}}}{n}=\bar{Y}-\hat{\beta}\bar{X} \quad \cdots (2\text{-}4)$$

(2-4)を(2-3)'に代入して、

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

$$= \bar{Y}\sum X_i - \hat{\beta}\bar{X}\sum X_i^2 + \hat{\beta}\sum X_i^2$$

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

$$\hat{\beta} = \frac{\sum X_i Y_i - \bar{Y}\overbrace{\sum X_i}^{n\bar{X}}}{\sum X_i^2 - \bar{X}\underbrace{\sum X_i}_{n\bar{X}}} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2}$$

$$= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \quad \cdots (2\text{-}5)$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \quad \cdots (2\text{-}4)$$

$$\hat{\beta} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \quad \cdots (2\text{-}5)$$

3　単純回帰モデル

回帰モデルとして、単純回帰モデル　$Y_i = \alpha + \beta Y_i + u_i$　を考える。

これは、Y_i（例えば消費）を X_i（例えば所得）のみで完全に説明できる場合のモデル（決定的モデル）　$Y_i = \alpha + \beta Y_i$　に u_i（誤差項）を追加したモデルであり、u_i を確率変数と仮定する確率的モデルである。

（1）単純回帰モデルの標準的仮定

仮定1　X_i は固定された値をとる（確率変数ではない）。

仮定2　$n \to \infty$ の時、$\sum (X_i - \bar{X})^2 \to \infty$

仮定3　$E[u_i] = 0$　（すべての i について）

仮定4　$V[u_i] = E[u_i^2] = \sigma^2$　（すべての i について）

仮定5　$Cov[u_i, u_j] = 0$　（すべての $i \neq j$ について）

仮定6　$u_i \sim N(0, \sigma^2)$（すべての i について u_i は平均 0、分散 σ^2 の正規分布に従う。）

（２）単純回帰モデルの係数：$\hat{\alpha}$、$\hat{\beta}$

①$\hat{\alpha}$、$\hat{\beta}$の確率的表現

$$\hat{\beta}=\frac{(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2} \quad \cdots (2\text{-}5)$$ この式（Y_iの関数）を確率変数u_iの関数に置き換える。

$Y_i=\alpha+\beta X_i+u_i$ 、$\bar{Y}=\alpha+\beta\bar{X}+\bar{u}$ より

$$Y_i-\bar{Y}=\beta(X_i-\bar{X})+u_i-\bar{u} \quad \cdots (3\text{-}1)$$

(3-1)を(2-5)に代入して、

$$\hat{\beta}=\frac{\sum(X_i-\bar{X})\{\beta(X_i-\bar{X})+u_i-\bar{u}\}}{\sum(X_i-\bar{X})^2}=\frac{\beta\sum(X_i-\bar{X})^2+\sum(X_i-\bar{X})(u_i-\bar{u})}{\sum(X_i-\bar{X})^2}$$

$$=\beta+\frac{\sum(X_i-\bar{X})(u_i-\bar{u})}{\sum(X_i-\bar{X})^2}=\beta+\frac{\sum\{(X_i-\bar{X})u_i-(X_i-\bar{X})\bar{u}\}}{\sum(X_i-\bar{X})^2}$$

$$=\beta+\frac{\sum(X_i-\bar{X})u_i-\bar{u}\overbrace{\sum(X_i-\bar{X})}^{0}}{\sum(X_i-\bar{X})^2}=\beta+\frac{\sum(X_i-\bar{X})u_i}{\sum(X_i-\bar{X})^2} \quad \cdots (3\text{-}2)$$

また、

$$\hat{\alpha}=\bar{Y}-\hat{\beta}\bar{X}=(\alpha+\beta\bar{X}+\bar{u})-\hat{\beta}\bar{X}$$

$$=\alpha-(\hat{\beta}-\beta)\bar{X}+\bar{u} \quad \cdots (3\text{-}3)$$

②$\hat{\alpha}$ 、$\hat{\beta}$の期待値

$\hat{\beta}$の期待値：$E[\hat{\beta}]=E\left[\beta+\frac{\sum(X_i-\bar{X})u_i}{\sum(X_i-\bar{X})^2}\right]=E[\beta]+E\left[\frac{\sum(X_i-\bar{X})u_i}{\sum(X_i-\bar{X})^2}\right]$

$$=\beta+\frac{\sum(X_i-\bar{X})\cdot\overbrace{E[u_i]}^{0}}{\sum(X_i-\bar{X})^2}$$

$=\beta$ 　　$\therefore \hat{\beta}$はβの不偏推定量である。

$\hat{\alpha}$ の期待値： $E[\hat{\alpha}]=\alpha-\bar{X}\cdot\underbrace{E[\hat{\beta}-\beta]}_{0}+\underbrace{E[\bar{u}]}_{0}$

$=\alpha$　　$\therefore\hat{\alpha}$ は α の不偏推定量である。

③ $\hat{\alpha}$、$\hat{\beta}$ の分散と共分散

$\hat{\beta}$ の分散： $V[\hat{\beta}]=E[\hat{\beta}-\beta]^2=E\left[\dfrac{\sum(X_i-\bar{X})u_i}{\sum(X_i-\bar{X})^2}\right]^2$

$$=E\left[\frac{\sum(X_i-\bar{X})^2u_i^2+\sum_i\sum_j(X_i-\bar{X})(X_j-\bar{X})u_iu_j}{\left\{\sum(X_i-\bar{X})^2\right\}^2}\right]$$

$$=E\left[\frac{\sum(X_i-\bar{X})^2u_i^2}{\left\{\sum(X_i-\bar{X})^2\right\}^2}\right]+E\left[\frac{\sum_i\sum_j(X_i-\bar{X})(X_j-\bar{X})u_iu_j}{\left\{\sum(X_i-\bar{X})^2\right\}^2}\right]$$

$$=\frac{\sum(X_i-\bar{X})^2\cdot\overbrace{E[u_i^2]}^{\sigma^2}}{\left\{\sum(X_i-\bar{X})^2\right\}^2}+\frac{\sum_i\sum_j(X_i-\bar{X})(X_j-\bar{X})\cdot\overbrace{E[u_iu_j]}^{0}}{\left\{\sum(X_i-\bar{X})^2\right\}^2}$$

$$=\frac{\sigma^2}{\sum(X_i-\bar{X})^2}\quad\cdots\quad(3\text{-}4)$$

〈凡人のメモ〉

私は、上記 ○ 部分の式変形がなかなか理解できませんでしたが、以下の具体例で何とか納得することができました。突然 j が出てくるのも納得しがたいところです。２つのΣ(シグマ)を区別する必要があるので仕方ないのでしょうが、何とかならないものか。

$$\left(\sum_{i=1}^{3}a_ib_i\right)^2=(a_1b_1+a_2b_2+a_3b_3)^2$$

$$=\underbrace{a_1^2b_1^2+a_2^2b_2^2+a_3^2b_3^2}_{i=j\text{の場合}}+\underbrace{2a_1a_2b_1b_2+2a_2a_3b_2b_3+2a_3a_1b_3b_1}_{i\neq j\text{の場合}}=\underbrace{\sum_{i=1}^{3}a_i^2b_i^2}_{i=j\text{の場合}}+\underbrace{\sum_{i=1}^{3}\sum_{j=1}^{3}a_ia_jb_ib_j}_{i\neq j\text{の場合}}$$

$\hat{\alpha}$ の分散：$V[\hat{\alpha}]=E[\hat{\alpha}-\alpha]^2=E\left[-\left(\hat{\beta}-\beta\right)\bar{X}+\bar{u}\right]^2$

$$=E\left[\left(\hat{\beta}-\beta\right)^2\bar{X}^2-2\left(\hat{\beta}-\beta\right)\bar{X}\bar{u}+\bar{u}^2\right]$$

$$=\bar{X}^2\cdot E\left[\left(\hat{\beta}-\beta\right)^2\right]-2\bar{X}\cdot\underbrace{E\left[\left(\hat{\beta}-\beta\right)\bar{u}\right]}_{0(※1)}+E\left[\bar{u}^2\right]$$

$$=\bar{X}^2\cdot V\left[\hat{\beta}\right]-0+\frac{\sigma^2}{n}=\bar{X}^2\cdot\frac{\sigma^2}{\sum\left(X_i-\bar{X}\right)^2}+\frac{\sigma^2}{n}$$

$$=\sigma^2\left\{\frac{\bar{X}^2}{\sum\left(X_i-\bar{X}\right)^2}+\frac{1}{n}\right\}$$

$$=\sigma^2\left\{\frac{\overbrace{n\bar{X}^2+\sum\left(X_i-\bar{X}\right)^2}^{\sum X_i^2(※2)}}{n\sum\left(X_i-\bar{X}\right)^2}\right\}$$

$$=\frac{\sigma^2\sum X_i^2}{n\sum\left(X_i-\bar{X}\right)^2}\quad\cdots(3\text{-}5)$$

(※1) $E\left[\left(\hat{\beta}-\beta\right)\bar{u}\right]=E\left[\frac{\sum\left(X_i-\bar{X}\right)u_i}{\sum\left(X_i-\bar{X}\right)^2}\cdot\frac{1}{n}\sum u_i\right]=\frac{E\left[\sum_i\left(X_i-\bar{X}\right)u_i\sum_j u_j\right]}{n\sum\left(X_i-\bar{X}\right)^2}$

$$=\frac{\sum_i\sum_j\left(X_i-\bar{X}\right)\cdot\overbrace{E\left[u_iu_j\right]}^{\sigma^2}}{n\sum\left(X_i-\bar{X}\right)^2}=\frac{\overbrace{\sum_i\left(X_i-\bar{X}\right)\sigma^2}^{0}}{n\sum\left(X_i-\bar{X}\right)^2}=0$$

(※2) $n\bar{X}^2+\sum\left(X_i-\bar{X}\right)^2=n\bar{X}^2+\sum\left(X_i^2-2X_i\bar{X}+\bar{X}^2\right)$

$$=n\bar{X}^2+\sum X_i^2-2\underbrace{\sum X_i}_{n\bar{X}}\bar{X}+\underbrace{\sum\bar{X}^2}_{n\bar{X}^2}=\sum X_i^2$$

$\hat{\alpha}$、$\hat{\beta}$ の共分散：$Cov[\hat{\alpha},\hat{\beta}]=E[(\hat{\alpha}-\alpha)(\hat{\beta}-\beta)]=E[\{-(\hat{\beta}-\beta)\bar{X}+\bar{u}\}(\hat{\beta}-\beta)]$

$$=-E[\hat{\beta}-\beta]^2\bar{X}+\underbrace{E[\bar{u}(\hat{\beta}-\beta)]}_{0}$$

$V[\hat{\beta}]=E[\hat{\beta}-\beta]^2=\dfrac{\sigma^2}{\sum(X_i-\bar{X})}$ より、

$$Cov[\hat{\alpha},\hat{\beta}]=-\frac{\sigma^2\bar{X}}{\sum(X_i-\bar{X})^2}=-\frac{\sigma^2\bar{X}}{\sum X_i^2-n\bar{X}^2} \quad \cdots (3\text{-}6)$$

④σ^2 の不偏推定量：s^2

$s^2=\dfrac{\sum\hat{u}_i^2}{n-2}$ は、σ^2 の不偏推定量である。

（証明）

$\hat{u}_i=Y_i-\hat{\alpha}-\hat{\beta}X_i=\alpha+\beta X_i+u_i=u_i-(\hat{\alpha}-\alpha)-(\hat{\beta}-\beta)X$

$\therefore u_i=(\hat{\alpha}-\alpha)+(\hat{\beta}-\beta)X_i+u_{ii}$

両辺を２乗して、

$u_i^2=(\hat{\alpha}-\alpha)^2+(\hat{\beta}-\beta)^2X_i^2+\hat{u}_i^2+2(\hat{\alpha}-\alpha)(\hat{\beta}-\beta)X_i+2(\hat{\alpha}-\alpha)\hat{u}_i+2(\hat{\beta}-\beta)X_i\hat{u}_i$

$\sum u_i^2=n(\hat{\alpha}-\alpha)^2+(\hat{\beta}-\beta)^2\sum X_i^2+\sum\hat{u}_i^2+2(\hat{\alpha}-\alpha)(\hat{\beta}-\beta)\underbrace{\sum X_i}_{n\bar{X}}+2(\hat{\alpha}-\alpha)\underbrace{\sum\hat{u}_i}_{0}+2(\hat{\beta}-\beta)\underbrace{\sum X_i\hat{u}_i}_{0}$

$=n(\hat{\alpha}-\alpha)^2+(\hat{\beta}-\beta)^2\sum X_i^2+\sum\hat{u}_i^2+2n\bar{X}(\hat{\alpha}-\alpha)(\hat{\beta}-\beta)$

$E[\sum u_i^2]=n\underbrace{E[\hat{\alpha}-\alpha]^2}_{\hat{\alpha}\text{の分散}}+\sum X_i^2\underbrace{E[\hat{\beta}-\beta]}_{\hat{\beta}\text{の分散}}+E[\sum\hat{u}_i^2]+2n\bar{X}\cdot\underbrace{E[(\hat{\alpha}-\alpha)(\hat{\beta}-\beta)]}_{\hat{\alpha},\hat{\beta}\text{の共分散}}$

$n\sigma^2=n\left\{\dfrac{\sigma^2\sum X_i^2}{n\sum(X_i-\bar{X})^2}\right\}+\sum X_i^2\dfrac{\sigma^2}{\sum(X_i-\bar{X})^2}+E[\sum\hat{u}_i^2]+2n\bar{X}\left(-\dfrac{\sigma^2\bar{X}}{\sum(X_i-\bar{X})^2}\right)$

$=2\sigma^2\underbrace{\left\{\dfrac{\sum X_i^2-n\bar{X}^2}{\sum(X_i-\bar{X})^2}\right\}}_{1}+E[\sum\hat{u}^2]$

$=2\sigma^2+E[\sum\hat{u}^2]$

$(n-2)\sigma^2=E[\sum\hat{u}^2]$

$E\left[\dfrac{\sum\hat{u}^2}{n-2}\right]=E[s^2]=\sigma^2$　　$\therefore s^2$はσ^2の不偏推定量である。

⑤$\hat{\alpha}$、$\hat{\beta}$の不偏分散

(3-5)および(3-4)のσ^2をそれぞれs^2に置き換えると、

$$s_{\hat{\alpha}}^2=\frac{s^2\sum X_i^2}{n\sum\left(X_i-\bar{X}\right)^2} \quad \therefore s_{\hat{\alpha}}^2は\hat{\alpha}の分散\left(\sigma_{\hat{\alpha}}^2\right)の不偏推定量である。$$

$$s_{\hat{\beta}}^2=\frac{s^2}{\sum\left(X_i-\bar{X}\right)^2} \quad \therefore s_{\hat{\beta}}^2は\hat{\beta}の分散\left(\sigma_{\hat{\beta}}^2\right)の不偏推定量である。$$

（３）単純回帰モデルの仮説検定：t検定

①$\hat{\alpha}$、$\hat{\beta}$の分布

(3-2)より、

$$\hat{\beta}=\beta+\frac{\sum\left(X_i-\bar{X}\right)u_i}{\sum\left(X_i-\bar{X}\right)^2}=\beta+\sum\left\{\frac{X_i-\bar{X}}{\sum\left(X_i-\bar{X}\right)^2}\right\}u_i$$

これを(3-3)に代入して

$$\hat{\alpha}=\alpha-\bar{X}\frac{\sum\left(X_i-\bar{X}\right)u_i}{\sum\left(X_i-\bar{X}\right)^2}+\underbrace{\frac{1}{n}\sum u_i}_{\bar{u}}=\alpha-\sum\left\{\frac{\bar{X}\left(X_i-\bar{X}\right)}{\sum\left(X_i-\bar{X}\right)^2}-\frac{1}{n}\right\}u_i$$

$u_i\sim N\left(0,\sigma^2\right)$（仮定６）より、$\hat{\alpha}$、$\hat{\beta}$も正規分布に従う。

②t分布の導出（$\hat{\alpha}$については省略）

$\hat{\beta}$は正規分布に従うことから、これを標準化すると、

$$Z=\frac{\hat{\beta}-\beta}{\sigma_{\hat{\beta}}}=\frac{\hat{\beta}-\beta}{\frac{\sigma}{\sqrt{\sum\left(X_i-\bar{X}\right)^2}}} \quad \sim N(0,1) \quad ・・・(3\text{-}7)$$

(3-7)のσをsで置き換えると、

$$t=\frac{\hat{\beta}-\beta}{s_{\hat{\beta}}}=\frac{\hat{\beta}-\beta}{\frac{s}{\sqrt{\sum\left(X_i-\bar{X}\right)^2}}} \quad \sim t(n-2) \quad ・・・(3\text{-}8)$$

(3-8)を以下のように変形する。

$$t=\frac{\hat{\beta}-\beta}{s_{\hat{\beta}}}=\frac{\hat{\beta}-\beta}{\frac{s}{\sqrt{\sum\left(X_i-\bar{X}\right)^2}}}=\frac{\frac{\hat{\beta}-\beta}{\frac{\sigma}{\sqrt{\sum\left(X_i-\bar{X}\right)^2}}}}{\frac{\frac{s}{\sqrt{\sum\left(X_i-\bar{X}\right)^2}}}{\frac{\sigma}{\sqrt{\sum\left(X_i-\bar{X}\right)^2}}}}=\frac{\frac{\hat{\beta}-\beta}{\sigma_{\hat{\beta}}}}{\frac{s}{\sigma}}=\frac{\frac{\hat{\beta}-\beta}{\sigma_{\hat{\beta}}}}{\sqrt{\frac{\frac{(n-2)s^2}{\sigma^2}}{n-2}}}=\frac{Z}{\sqrt{\frac{Y}{n-2}}}$$

〈凡人のメモ〉

残差分散 $\frac{\sum u_i^2}{n-2}$ を s^2 と表記している参考書もありますが、私は s^2 が出てくると無意識のうちに $\frac{\sum\left(X_i-\bar{X}\right)^2}{n-1}$ として考えていたため、回帰モデルの t 分布がなかなか理解できませんでした。残差分散を $\hat{\sigma}^2$ として区別している参考書もありますが、このサブノートではどちらも s^2 と表記しています。