講 座

使える統計学基礎講座 (第3回)

—直線回帰と分散分析—

光永貴之

独立行政法人農業・食品産業技術総合研究機構中央農業総合研究センター

Lectures of Statistical Logics for Beginners (Part III)

Linear Regression and Analysis of Variance

Takayuki MITSUNAGA

National Agricultural Research Center NARO

1. はじめに

第二回では、相関係数と回帰分析について実用上の問題点を中心にお話ししました。今回は前回お話ししきれなかった直線回帰の検定と解釈の注意点について最初に述べていきます。続いて、最も重要な検定手法である分散分析の原理等について述べていきます。

第一回でもお話ししました様に、この講座はできる限り数学をいじらない様に心がけています。数学的な計算過程、証明に興味のある方は藤井(1983)などを参考にすることをお勧めします。また、分散分析やその応用である実験計画法については既に本誌上で講座として優れた解説が行われています(執行, 2002)。従って本稿では、データを取り扱う際の注意点や解析上の問題点等を中心にお話ししていきます。試験設計、論文作成などの際の参考にして頂ければ幸いです。

2. 直線回帰の検定

前回も述べました様に、回帰直線を計算した際には得られた直線について少なくとも二つの検定を行う必要があります。第一は得られた直線の傾

きが0(傾きなし)と統計的に有意に違うかどうかの確認です。第二に、得られた直線の切片が0と有意に違うかの確認です。もしも、回帰直線の計算にJMPなどの統計ソフトを用いていれば、おそらく自動的に検定をかけてくれているはずです。例えば、JMPで回帰直線を求めた場合、図1のような結果が出力されます。用いたデータは前回の相関係数の解説で使ったウサギの体重と食べた草の量の関係を表したものです(メスのデータ)。図1下側の丸で囲まれた部分が傾きと切片の検定の部分です。統計ソフトを用いて直線回帰を行う場合には最初ここを見て得られた直線のパラメーターが0と有意に異なることを確認する必要があります。特に、傾きが0と有意でない場合には直線回帰そのものが意味のないものになってしまいます。ちなみに、図1の丸で示した部分の上にある“分散分析”とある所も同じことを別のやり方で表しているものです。分散分析については後述するので詳しいことはここでは述べませんが、この項が有意である場合には直線の傾きが0と有意に異なることを表しています。統計ソフトを使わずに回帰直線を計算した場合には、これらの検定は自分で行う必要がありますが、手計算で行う場合についてはここでは触れません。それほど難しい計算ではありませんので藤井(1983)などを参照してください。

二つの回帰係数が異なるかどうかについても検定をかけることもできます。手計算をする場合、この検定は前処理としての計算が結構大変です

平成22年5月10日受付

平成22年5月14日受理

Corresponding author

光永貴之 Takayuki MITSUNAGA

〒305-8666 茨城県つくば市観音台3-1-1

3-3-1 Kannondai, Tsukuba, Ibaraki, 305-8666, Japan

E-mail: aeiou@affrc.go.jp

体重と食べた草の量の二変量の関係 雌雄=メス

直線のあてはめ

直線のあてはめ

食べた草の量 = 8.7099099 + 8.1467181*体重

あてはめの要約

R2乗 0.584812
自由度調整R2乗 0.569984
誤差の標準偏差(RMSE) 1.476226
Yの平均 22.96667
オブザベーション(または重みの合計) 30

分散分析

要因	自由度	平方和	平均平方	F値
モデル	1	85.94788	85.9479	39.4393
誤差	28	61.01879	2.1792	p値(Prob>F)
全体(修正済み)	29	146.96667		<.0001*

パラメータ推定値

項	推定値	標準誤差	t値	p値(Prob> t)
切片	8.7099099	2.2861	3.81	0.0007*
体重	8.1467181	1.297233	6.28	<.0001*

図1 統計ソフト JMP ver8.0 を用いて直線回帰を行った例。用いたデータは第二回で紹介したウサギの体重と食べた草の量の関係のもの(メスのデータ)。直線回帰を行った場合、最初に確認することはパラメーター推定値の有意性である(図中の丸で囲まれた部分)。

が、ここでは触れません。藤井(1983)などを参照してください。統計ソフトを用いる場合、いくつかのソフトでは一見すると対応する検定がないように思える場合があります。例えば、ここでよく紹介する JMP では、回帰の所をいくら探しても二直線間の比較の仕方を見つけることはできないと思います。その理由は同等の作業をもっと合理的に“共分散分析”という別の統計手法で行うからです。共分散分析についてはこの講座の最後で解説する予定ですが、ここではなぜこれを用いた方が合理的なのかだけを簡単に紹介しておきます。例として、前回用いたウサギと餌の量の関係を考えてみましょう。オスとメスのグループでそれぞれ体重と餌の量の関係を測りました(図2)。前回のお話では、オスとメスでは平均体重に違いがあるので別々にプロットしたと述べました。実際には、本当にオスとメスの平均体重が違うのかどう

a)

b)

図2 ウサギの体重と一日に食べる草の量の散布図(仮想データ)。前回の図6と同じ。雌雄で体重が目で見ただけで二つに分けた。どちらも有意な相関が見られる。
a) オスの場合。
b) メスの場合。

かについて、最初に何らかの統計検定をかけてジャッジする必要があります。従って、個別に回帰直線を求め、それから回帰係数の比較を行う場合には二段階の統計を行うことになります。ところで、統計検定とは自分の行った比較について“絶対に違う”ということを保証するものではありません。“5%水準で統計的に有意に異なる”とは、得られたデータから推測して比較した集団が同じである確率が5%以下であると言っているだけです。言い換えれば5%水準の判断とは、本当は比較した集団が同じであるのに違うとミスジャッジしてしまう確率(第一種の過誤の確率と言います)が5%であるということです。それならば、もっと基準を厳しくすればミスジャッジが減らせるかということ、判断水準を厳しくすればするほど本当は比較した集団が違うのに同じとミスジャッジしてしまう確率(第二種の過誤の確率)が増加

してしまいます。結局の所、ほとんどの場合に用いられている“5%水準で有意性を判断する”という基準は、第一種の過誤の確率を0.05までは認めましょう、というものです。

さて、このことを踏まえて先ほどの例を考えてみましょう。最初にウサギの雌雄の違いを検定して（一回目の統計）、次に回帰係数を比較する（二回目の統計）場合、ジャッジは2回行っています。一回につき、5%の間違いの確率がありますから二回とも間違えない確率は $(1-0.05)(1-0.05)=0.9025$ となり、全体ではおよそ1割間違えてしまうことになります。このように統計判断をシークエンシャルに繰り返すことによって全体として判断を誤る確率が増加することを“多重性の問題”と呼びます。もう少しこれについて知りたい方は光永（2009）などを参考にしてください。一方、共分散分析は要因間（雌雄）の比較と傾きの比較を後述の分散分析と同等の方法で同時に行う手法ですので、多重性の問題が解消されます。以上の理由から、計算された回帰直線間の傾きの比較は安易に行うものではないということが理解できます。

3. 直線回帰とデータに対する適合性

直線回帰の目的は線形関係にあることが予想される二変数について、説明変数の大きさに応じた目的変数の大きさを予測することです。つまり、“良い回帰”というのは、ある大きさの説明変数に対して精度よく目的変数を推定できるものであることが分かります。

直線回帰の精度の良さは得られた回帰直線と実際のデータのばらつき具合で決まります（図3）。もしも、データが一直線上に分布していれば（つまり、相関係数=1の状態ですね）、もちろん全てのデータは回帰直線上に乗ることになります。このような場合には、得られた回帰直線によってほぼ完全に目的変数を予測できるはずですが（図3a）。一方、データがxy平面上にばらついた状態になっている時には（つまり、相関係数が小さい場合ですね）回帰直線はデータのばらつきの中を通るような直線になります。このような場合、得られた回帰直線から推定される目的変数の精度はあまり高くはならないと考えられます（図3b）。データのばらつきが大きければ大きいほど、言い換えれば相関係数が小さければ小さいほど回帰直

図3 データのばらつきと直線回帰の関係。
a) データが一直線上にある場合（相関係数=1）。得られた回帰直線は目的変数を完全に予測できる。
b) データがばらついている場合（相関係数<1）。得られた回帰直線の推定する目的変数は実際のデータとおそらく一致しない。

線の予測精度は減少していくことが予想されます。これらのことを考えると直線回帰の予測精度とデータの相関係数の間には何らかの関係がありそうです。

この関係を端的に表しているものが“決定係数”と呼ばれるものです。決定係数は r^2 （または R^2 ）で表されますが、これは積率相関係数の2乗であるということです。決定係数が“決定”しているものは、目的変数の平均値からの分散を回帰直線の傾きが何%説明しているのかということです。従って、決定係数が大きければ大きいほど実際のデータのばらつきを回帰直線がよく表していることになり、結果として回帰直線の予測精度は高いと言えます。決定係数、相関係数、回帰係数の関係についての数学的な証明は省略します。藤井（1983）などを参照してください。ここでは決定係数の捉え方について取り上げましょう。例として先ほど示した図1の中段部分の“あてはめの要約”という所を見てみます。決定係数は0.584812と

なっていますから、目的変数（食べた餌の量）の平均からのばらつきは回帰直線によっておよそ6割弱が説明されていることになります。それならば、残りの4割は何かというと実験誤差や個体差等による回帰では説明できないばらつきです。一般に化学反応等のデータでは決定係数は極めて高い値を示しますが、生物個体等のデータではそれほど高い値は示しません。その理由はどうしても個体差が大きく影響してしまうからです。一概には言えませんが、大体5～7割以上のばらつきが説明できればそれなりに使える回帰と言えます。決定係数はあくまでも指標であり、単に値が大きいことを見るだけではなく、残りの分散がどのような性質であるかを考えることが重要です。例えば、化学反応のデータで決定係数が0.8であれば、データの分散の20%は何らかの実験誤差ですから、決してコントロールされた実験とは言えません。逆に、野外の調査で決定係数が0.7であった場合、気温や降水量、日照等のコントロールできない要因の影響が30%しかばらつきを作らない訳ですから良いデータと言えます。

前回、相関係数の有意性はサンプル数に依存するというお話をしました。この性質は決定係数にも受け継がれます。また、説明変数を複数考える重回帰分析の場合、説明変数が多くなるほど見かけ上の決定係数は大きくなります。このような問題を補正するために“自由度調整済み決定係数”というものを用いられる場合があります。これは、決定係数を回帰に用いている説明変数やデータ数で重み付けしたものです。一般には、元の決定係数に比べて、自由度調整済み決定係数の方が少し小さくなりますが、特にデータ数の違う回帰式を見比べる際に便利です。

直線回帰によって推定された目的変数がどのくらいの誤差を持つのかということは回帰直線の95%信頼限界で表すのが一般的です。95%信頼限界については第一回でお話しした様に、推定された値が95%の確率で収まる範囲を示します。例として先ほど示した図1の上段部分のグラフを見てみましょう。回帰直線の周りにある破線がこの回帰直線の95%信頼限界を示しています。この図を見て気づくのは中心、つまり説明変数と目的変数の平均値の近くでは95%信頼限界は狭いのですが、両端に行くにつれてだんだんと広がっていることです。これは前回に示した回帰直

線の式が基本的に両変数の平均からのずれを評価していることと関係があります。もう一度回帰直線の式を示しておきます。

$$Y - Y_M = \frac{Cov(X, Y)}{Var(X)}(X - X_M) \dots\dots\dots(1)$$

この式を考えると回帰直線は必ず説明変数の平均値と目的変数の平均値を通ることが分かります。なぜならば、どちらかの平均値を式に代入すれば、もう一方の変数の値は平均値になるからです。つまり、回帰直線の“中心”はデータの平均値によって決まり、データのばらつきによる直線のずれは傾きに影響を与えることが予想されます。傾きがずれた場合、直線の“中心”付近ではそれほど大きな影響が出ませんが、中心から離れて行くにつれてずれはどんどん大きくなります。これが、回帰直線の信頼限界が図1上段のように歪んだ形になる理由です。このことは直線回帰の予測精度はデータ分布の中心から離れるほど減少していくことを示します。

回帰直線の信頼限界の性質を理解しておくことは実験計画の決定や得られた回帰直線の利用の際に大変重要です。第一に中心から大きく外れた場所の予測は誤差が大きく信頼性が低いことが挙げられます。しかし、通常は予測したい部分こそ説明変数のデータを取らなかった外側の場所なのでこれは大きなジレンマです。

第二に、得られた回帰直線から説明変数の逆推定を行う場合、その推定値の誤差は上下対称にはならないことが挙げられます。逆推定とは、得られた回帰直線を用いて、ある目的変数を得るための説明変数を推定することです。例えば、図4を見てみましょう。これは図1の上段部分の拡大です。仮に一日に25g食べるメスウサギの体重を逆推定すると、2kgであることが分かります。ただし、信頼限界は下側が1.9kgに対して上側が2.13kgと上下で幅が異なります。

このような信頼限界の歪みは説明変数の分布に大きく影響されます。例として図5を見てみましょう。これは説明変数の幅と95%信頼限界の形を比較したものです。図5aは説明変数に10, 15, 20の3点を取り、それぞれで10反復取った場合です。一方、図5bは説明変数の幅を広げて5, 10, 15, 20, 25, 30の6点を取り、それぞれで5反復取った場合です。両者の破線で示された95%信

体重と食べた草の量の二変数の関係 雌雄=メス

図 4 回帰直線と逆推定の例. 食べた草の重さが 25g の場合のウサギ (メス) の体重を逆推定する. 推定された体重は 2 kg だが, その信頼限界は下側が 1.9 kg に対して上側が 2.13 kg となり, 上下の幅が違うことに注意.

信頼限界を見てみると明らかに図 5b の方が狭くなっています. 特に, データの中心から離れた場所では顕著です. このことから, 直線回帰の推定精度を上げるためにはなるべく説明変数の幅を広く取るようにすることが重要であることが分かります.

4. 直線回帰とデータ操作による問題

直線回帰の精度は決定係数と 95% 信頼限界によって示されるとお話ししました. しかし, これらの指標を読み手として判断するには注意が必要です. 先ほど述べました様に, 決定係数の値そのものを重視するのではなくその背景を考察することはもちろん, それ以外にも落とし穴がある場合があります. 特に学会発表や論文で比較的良く見られる, データ操作によって見かけ上極めて高い決定係数が示される場合には要注意です.

例として, 図 6 に示した例を考えてみましょう. このデータは栽培温度によってエノキダケの菌糸の成長がどのくらい決定されるのかを調べたものです (仮想データ). 図 6a は測定したデータをそのまま直線回帰したものです. 一見すると分かりますが, 確かに栽培温度が上がるにつれて菌糸が伸びている様に見えます. しかし, それ以外の要因も影響しているようで決定係数は 0.5679

a) 説明変数と目的変数の二変数の関係 実験=グループ 1

b) 説明変数と目的変数の二変数の関係 実験=グループ 2

図 5 説明変数の分布によって 95% 信頼限界の形が変わる例. どちらのグラフも回帰直線は $y=x$. サンプルサイズは 30.

- a) 説明変数 10, 15, 20 に 10 サンプルずつ取った場合.
- b) 説明変数 5, 10, 15, 20, 25, 30 に 5 サンプルずつ取った場合.

とそれほど高くありません. 次に全く同じデータをそれぞれの温度区で平均値を取ってから直線回帰を行ってみましょう (図 6b). 驚くべきことに生データで回帰を行った場合と全く同じ回帰直線が得られた上に, 決定係数が 0.9577 と恐ろしく高くなっています. 図示していませんが, 95% 信頼限界も大変小さくなります. なぜなら決定係数が大きいということは, 直線の傾きで目的変数のばらつきがほとんど説明されてしまい, それ以外のばらつきはほとんどないからです. 一見すると図 6a よりも図 6b の方が優れた結果に見えますが, 本当にそうなのでしょうか?

第一回でもお話しした様に, 平均値を取るという作業は, データ集団の性質を感覚的に捉えやすくするためのものです. しかし, 確かに分か

図6 エノキダケの栽培温度と一日に伸びる菌糸の長さの関係（仮想データ）。

- a) データをそのまま直線回帰した場合、回帰直線は $y = 1.02x + 9.5$ 、決定係数は 0.5679 とそれほど高くない。
b) それぞれの温度区で平均を取ってから直線回帰した場合、回帰直線は $y = 1.02x + 9.5$ 、決定係数は 0.9577 と極めて高い。

りやすくはなりますが、同時に“データがどれだけばらついているか？”という重要な情報が失われてしまいます。これが図6bで見かけ上決定係数が大きくなっている理由です。さて、直線回帰の目的は、“ある説明変数の大きさの時、予想される目的変数はどれくらいになるか？”ということを“対象となる現象の性質に応じて”精度よく推定することです。従って、問題としている現象がばらつきやすいものならば、そのばらつきを考慮して幅を持たせた推定になることは本来受け入れるべきものです。図6aから導かれる推定値は実際のデータのばらつきを考慮した95%信頼限界を提示するので問題はありません。一方、図6bから導かれる推定値は、それぞれの処理区の“平均値”の違いを考慮しているに過ぎません。従って、図6bから得られる95%信頼限界はあくまでも“平均値の95%信頼限界”であり、データのばら

つきは説明できません。

もっと話を膨らませてみましょう。図6を見たエノキダケ栽培業者さんにとって図6aと6bはどちらが有益な情報でしょうか？業者さんが実際の栽培に当たり、栽培温度を決定するために図6から逆推定をしたとします。図6aからの逆推定は実際のデータのばらつきを考慮するためによりかなり広い幅の推定値になります。この推定値から栽培温度を判断すれば失敗は少ないでしょう。一方、図6bから逆推定を行えば、温度の推定幅は大変小さくなります。これに従って栽培を行えばかなりの割合で期待通りのエノキの成長が得られないと考えられます。これでは回帰の意味がありません。

回帰に限らず、統計を行う際に平均値を取ってから解析を行うという手順はあまり勧められるものではありません。確かに見かけ上美しくなりますが、現実が美しくなる訳ではありません。データ解析において目指すべきなのは、“取ったデータを美しくする”ことではなく、“現実のばらつきを受け入れ、解釈し、可能ならばコントロールする”ことです。直線回帰においては、十分な反復、広く取った説明変数の幅、そして精度の高い説明変数の測定になります。これらが満たされているならば、例えば決定係数が多少低くても得られた回帰直線は現象を理解する手助けになるはずです。

5. 分散分析の原理

分散分析はR.A. フィッシャーによって1920年代に開発された統計手法です。最近では概念をより拡張させた“一般化線形モデル”に主役の座を譲った感がありますが、現在でも最も重要な統計手法の一つであることは間違いありません。考案されてから80年にわたって主要なツールとして用いられてきた理由はその優れた構造性と論理性にあります。

分散分析は次の式を仮定することがスタートです。

$$Y = \mu + A_i + \varepsilon \dots\dots\dots(2)$$

ここでYは実際のデータ値、 μ はある現象の本来の値、 A_i はi番目の要因の効果による増減分、 ε は誤差による増減分を示します。この式は私たちが実際に測定したデータは、要因の効果や誤差の足し算で表されているということを仮定していま

す。もちろん、これはあくまでも仮定であり、現実がこの式に従っている必要はありません。しかし、問題としている要因が何らかの影響を示すならば、そこそこの近似式としては使えるはず（この部分は数学的には A を変数とした関数のテイラー展開として考えると分かりやすいかもしれませんが）。さて、(2)式を受け入れて、かつ μ , A , ε が互いに独立であると考えましょう。これは言うてしまえば、要因に関係なく誤差は影響しているという仮定ですので問題はないと思います。この時、“分散の線形性”という便利な性質から、次の式が成り立ちます。

$$V(Y) = V(\mu) + V(A) + V(\varepsilon) \dots\dots\dots(3)$$

この式の意味するものは、実際のデータの分散は、 μ の分散（これは実は 0 です。なぜならある現象の本来の値は決まっていると仮定できるからです。）と要因の効果の分散、そして誤差の分散の和であるということです。ただし、この式はあくまでも理論的な関係を表します。実際の分析においては、分散は推定値になりますのでこの等式は成り立ちません。

さて、もしもある要因（仮に A_1 としましょう）の効果が測定したデータに影響を与えているならば、要因間の分散は 0 よりも大きくなることが期待できます。なぜならば、 A_1 の効果が高いということは、それ以外の要因との差が大きいということだからです。従って (3) 式の $V(A)$ を 0 と比較して何らかの基準に従って $V(A) > 0$ を示せば要因の効果があると結論できます。これが分散分析の原理です。

問題は、実際に $V(A)$, $V(\varepsilon)$ をどのように推定するのかということです。そのために、(3) 式をさらに変形します。ただし、 $V(\mu)$ は 0 ですので省きましょう。

$$\Sigma(Y - GM)^2 = \Sigma n(MA_i - GM)^2 + \Sigma(Y - MA_i)^2 \dots(4)$$

ここで Y は実際のデータ値、 GM はデータ全体の総平均、 n は要因毎の反復数、 MA_i は i 番目の要因のデータの平均値を示します。最初の項はこれを平均すればデータ全体の分散 $V(Y)$ になります。このような形のものを“平方和”と呼びます。次の項は平均すれば数学的には要因間分散になることが期待されます。最後の項は平均すれば誤差分散になります。なぜ、(4) 式のように平方和で表

図 7 要因間平方和と要因内平方和の分離例。土壤にクロム、モリブデンを添加した際のイネの生育阻害を調査（仮想データ）。要因毎のデータの平方和（要因内平方和）が誤差分散を作り出す。一方、要因毎の平均値の平方和（要因間平方和）が要因間分散を評価する。

記し直すのかということ、分散の推定が多少厄介だからです。図 7 を例として考えてみましょう。これは、土壤に一定濃度のクロム、モリブデンを添加した際のイネの生育阻害を調べたものです（仮想データ）。通常、実験を行う際にはそれぞれの処理区毎の反復は同じにしますので、ここでも同じにしています。これは重要な点です。誤差は要因に独立にかかりますから、クロムのデータだけのばらつき、モリブデンのデータだけのばらつき、対照区だけのデータのばらつきはそれぞれ誤差によるものと見なすことができ、同じ大きさであることが期待されます。これを“等分散性の仮定”と呼びます。実際のデータでは多少ばらついてしまっていますが、あまり気にしなくて大丈夫です。ただし、露骨に違う場合には分散分析ができないので注意が必要です。

さて、実験全体の誤差分散 $V(\varepsilon)$ は理解しやすく、要因毎の平方和を足したものの平均で表されます。一方、要因間分散 $V(A)$ はちょっと複雑です。なぜなら、要因内の平均値を計算する時点で誤差の影響を受けているからです。結論を言いますと、要因間平方和は平均すると統計的には誤差分散と要因間分散の和になってしまうことが分かっています。つまり、理論上は可能でも、実際

には要因間分散 $V(A)$ は単独では分離できず、 $V(A) + V(\epsilon)$ の形でしか推定できないということです。幸いなことに分離できなくても評価は可能ですので問題はありません。理論的には $V(A) > 0$ を評価するべきが、実際には $V(A) + V(\epsilon) > V(\epsilon)$ を評価するというだけです。具体的なデータからの計算と解釈については藤井(1983)などを参照してください。ここで覚えておくべきことは、全データの平方和は要因間と誤差の二つの成分に分解ができるということです。

先ほど、各処理区の反復数をそろえることが重要であると述べました。このような実験デザインを“バランスドデザイン”と呼びます。この場合、処理区毎の誤差分散は等価に考えることができます。しかし、実際の実験では最初に反復数をそろえたからといって、全ての反復がデータとして使えるとは限りません。図7の場合でも、何らかの原因で一部のイネが枯れてしまうことも考えられます。このような場合には、誤差分散を計算する際に反復数に応じて重み付けをしてやる必要が出てきます。なぜなら、反復の少ない処理区の誤差分散の推定精度はどうしても悪くなるからです。幸いなことに、現在用いられている統計ソフトはまず間違いなく自動的に補正してくれますのでこの点は安心です。ただし、反復数をそろえた方がよいことは言うまでもありません。

分散分析の目的が $V(A) + V(\epsilon) > V(\epsilon)$ を示すことである以上、良い実験デザインというのは如何に $V(\epsilon)$ を精度よく推定するかということにかかってきます。これは、各処理区内で十分な反復を取り、かつなるべくばらつきを小さくすること、そして特定の処理区だけばらつきが大きくなるようにすることを意味します。より具体的には、用いる容器をそろえる、試料のロットをそろえる、測定する人を決める、測定する機器をそろえる、などです。このような細かい作業の積み重ねによって $V(\epsilon)$ の精度は変わってきます。従って、実験の際にはこれらのことを注意することが重要です。

さて、具体的にどのようにして $V(A)$ 、 $V(\epsilon)$ の大きさを比較するのでしょうか？このためには $\{V(A) + V(\epsilon)\} / V(\epsilon)$ という比を取って“F検定”というものを行います。これについては次回にお話をさせていただきます。

6. 第三回のまとめ

1) 直線回帰を行った場合、得られた回帰係数、切片それぞれについて0と有意に異なるかどうかを判断するために検定を行う必要がある。特に回帰係数が0と有意に異なると言えない時には直線回帰を行った意味がない。係数や切片の検定は手計算でも可能だが、統計ソフトが自動的にしてくれる。

複数の回帰直線のパラメーターが違うかどうかを判断する検定も可能だが、この場合多重性の問題が生じるので共分散分析を行った方が合理的である。

2) 直線回帰がデータに当てはまっている指標として決定係数がある。これは目的変数のばらつきが何%回帰式によって説明されているかを表す。決定係数はデータの積率相関係数の2乗で示される。積率相関係数は同じ値であっても計算に用いたデータ数によって有意性が変化したが、この性質は決定係数にも引き継がれるので、データ数や説明変数の数で重み付けされる自由度調整済み決定係数が用いられることがある。

得られた回帰直線の精度は95%信頼限界で示される。これは逆推定を行う場合にも大変重要であるが、データ分布の中心では幅が狭く、外側に行くにつれて広がっていく。良い実験デザインはなるべく説明変数の幅を広くとることである。

3) 決定係数は直線回帰の精度を判断する重要な指標であるが、単に数字を比べるだけでは誤った解釈をしてしまう場合がある。特に、データをそれぞれの処理区毎に平均してから直線回帰を行うと異常に高い決定係数が示されることに注意しなければならない。また、このような平均化されたデータの直線回帰は見た目はきれいでも、実際の予測には役に立たない場合が多い。

4) 分散分析は最も広く用いられている統計手法の一つである。その原理は、線形モデルの仮定、分散の線形性に基づくデータ分散の分割、そして分割された分散間の違いをF検定によって検出するというものである。分散分析を行う場合には、誤差分散の推定精度を上げるためにバランスドデザインで実験を設定すること、十分な反復を取ること、実験条件をそろえること等が望ましい。

今回は、直線回帰については実用上の解釈の仕

方や注意点を中心にお話しました。また、分散分析についてはその元になっている考え方を理解して頂くことを目的に説明しました。広く普及しているこれらの手法も背景を理解しなければ結果の解釈に大きな誤解が生じてしまう可能性があることを理解して頂ければ幸いです。次回は分散分析の続きと共分散分析についてお話します。

8. 引用文献

- 執行盛之（2002）遠くまで飛ぶ紙ひこうきを折る
一直交表による実験計画法の適用一，農作業
研究 37（2）；81-96.
- 藤井宏一（訳）（1983）：生物統計学（原本：Sokal,
R.R. and F.J. Rohlf (1961) Introduction to
biometry），共立出版，東京，pp.447.
- 光永貴之（2009）：生物統計処理方法の注意点，植
物防疫 63（6）；377-380.