論　　文

VBAで開発した重回帰分析ツールの都道府県別標準化死亡比・国民栄養調査等のデータへの適用の試み

Trial Involving the Application of VBA-Developed Multiple Regression Analysis Tools on Japanese Prefectural Standardized Mortality Ratios and National Nutrition Surveys

堀　田　裕　史
HORITA Hiroshi

１．はじめに

１．１.VBA によるユーザーの判断を重視した重回帰分析ツールの概要

筆者は以前、重回帰分析用ツールをExcel上のVBAで開発した[1)]。このツールの特徴は次の点である。Ｎ個の説明変数による重回帰分析は、一段階前の（Ｎ－1）個の説明変数による重回帰分析での（Ｎ－1）個の説明変数に、それら以外の多数の説明変数の候補の一覧を決定係数の高い順に並べたリストを作り、その中からユーザーが適切と判断するものを選択し、Ｎ番目の説明変数として採用し、分析を進めていくことである。

この方式は、多数の説明変数の候補がある場合の重回帰分析への適合がよいと考えて開発し、実際に都道府県別の自殺率を、他の157種の都道府県別のデータを使って重回帰分析し、ツールの有用性を示した。

１．２．都道府県別食物摂取データ及び標準化死亡比データ

食物や栄養は、都道府県毎に均等に摂取されているだろうか。実際には、食塩の摂取量（男性）は1日10.5g～16.5gまで幅があり、肉類の摂取量（男性）は1日64g～115gまで幅がある。また消費統計においても、都道府県毎に大きく異なる場合がある。これら食物・栄養摂取等の違いが健康状態と関係していないか、気になるところである。

死因では徳島県が14年間糖尿病の死亡率日本一を続けるなど、死亡原因もまた都道府県毎に意外に大きく異なっている。都道府県の年齢構成の違いに対しては、標準化死亡比を使えば、年齢構成に依存しない都道府県毎の差異をとらえることができる。

では、都道府県毎に異なる食物・栄養摂取を使って、これまた都道府県毎に異なる標準化死亡比を分析できないか興味がもたれる。自殺率の重回帰分析では、単回帰分析（これは相関係数による分析にやや近い）とは異なり、自殺に影響する複数要因に対し複数の説明変数を使い分析した。標準化死亡比の場合も、食物・栄養摂取データを用いて分析できると思われる。ただし分析に際しては、分析を進めるための留意点や条件に注意し、また重回帰分析結果からリスク等の指標に代替しうる指標はどのようにして得られ

ほりた　ひろし（食物栄養学科）

るかを明らかにしつつ、重回帰分析を進める必要があるであろう。

２．VBAによる重回帰分析ツール適用上の留意点と課題

２．１．回帰分析と地域相関研究

重回帰分析では、被説明変数（目的変数、基準変数ともいう）を複数の説明変数を使って線型結合で予測する。ここでは、説明変数・被説明変数とも都道府県別のデータを使っている。一方、疫学（Epidemiology）のうちの記述疫学に属する生態学的研究（Ecological Study）の地域相関研究（Correlational Study）では、地域間のデータ間の関連性、特に相関を調べることになる[2]。よって本報告の都道府県別のデータ解析のうちの単回帰分析に関する記述は、地域相関研究としてよく紹介される内容と似ている。しかしながら重回帰分析は、結果が直感的に理解しにくいので、筆者は単回帰分析に匹敵するくらいに、なるべく分かりやすく、結果を視覚イメージで捉えやすくするのが望ましいのではないかと思っている。

２．２．因果関係

重回帰分析では、被説明変数と説明変数は、それぞれ説明される側、説明する側のデータである。原因と結果の観点では、被説明変数は結果、説明変数は原因であるので、この被説明変数と説明変数の間に因果関係が本来必要である。

疫学では、有病率など発症の程度を表す量を、それの原因となる要因に曝露の程度を示す量との関係を調べるが、発症の程度と曝露要因の間に因果関係が必要となる。因果関係は、原因は結果の前に生起すること（時間的順序）、原因と結果が他の知識と適合しているか（蓋然性）、他の研究の結果と一致しているか等、数個の条件を充足が求められ、疫学の方では、重要な検討事項となっている。本稿でも因果関係には注意を払うが、あくまで地域相関研究の一種であり、厳密な因果関係に立脚しているのではなく、可能性を示唆するに過ぎないことを前提としている。[3] [4] [5]。

２．３．リスク分析

健康関係のデータ分析では、発症率やロジットをデータとして、リスク比やオッズ比及びその信頼区間を求めることが多い。しかし、今回使用する死亡原因データが標準化死亡比であり、発症率等と異なりリスク比などが使えない。死亡原因のデータ分析であり、因果関係の問題を除いても、既存のリスク分析の代替となるような仕組みを組み入れる必要がある。これについては、「5.議論」で詳しく述べることにする。

２．４．ＶＢＡによる重回帰分析プログラムの拡張の必要性

前回作成した、重回帰分析プログラムは、Excel 2007の特にグラフィック関係で動作しない箇所が多々あったが、Excel 2003／2007の双方でエラーが発生しないようにした。

また、グラフ形式も若干変更を加えたこと、説明変数が7変数の場合の重回帰分析も行えるようにしたこと、更にリスクに関しての代替指標の提示のためのプログラムを追加したことにより、プログラム量は増加した。VBAによる主要部分は950行、総コーディング量は1390行となった。

３．分析に使用する健康データ

３．１．都道府県別死因データ

厚生労働省のホームページに「都道府県別死因の分析結果について」と題し、平成13～15年の都道府県別の標準化死亡比（ＳＭＲ:Standardized Mortality Ratio）が記載されている。分析は平成15年のデータで行った。死因は以下のものである[6) 7)]。

1) 脳血管疾患(男性)　2) 脳血管疾患(女性)　3) 心疾患(男性)　4) 心疾患(女性)
5) 糖尿病(男性)　6) 糖尿病(女性)　7) 胃がん(男性)　8) 胃がん(女性)
9) 肺がん(男性)　10) 肺がん(女性)　11) 大腸がん(男性)　12) 大腸がん(女性)
13) 肝がん(男性)　14) 肝がん(女性)　15) 子宮がん(女性)　16) 乳がん(女性)
17) 前立腺がん(男性)　18) 肺炎(男性)　19) 肺炎(女性)

３．２．国民栄養調査に基づくデータ

ここでは国民栄養調査そのものではなく、平成7年から平成14年までの国民栄養調査のデータを元に都道府県別にまとめたデータが公開されているのでこれを利用する[8)]。都道府県別の全体（男女計）及び男女別の各種栄養素の平均摂取量と、全体（男女計）の年齢構成の影響を排した栄養素摂取量の標準化比がホームページで公開されている。男女別と全体のデータがあるが、例えば都道府県別の男性の死因データの説明目的には、全体（男女計）ではなく、男性の栄養摂取データを優先して用いることとする。

３．３．都道府県別家計支出データ

総務省統計局の家計調査のホームページから、分類「家計収支編・総世帯・詳細結果表・年次・2008年」を選択して現れる統計表のうち、「（品目分類）第11表　都市階級・地方・都道府県庁所在市別１世帯当たり年間の品目別支出金額（総世帯）」という統計表を採用した[9)]。都道府県庁所在地別・品目別の全世帯年間支出平均額であり、円単位である。世帯人員で割り一人当たりの金額に換算し、都道府県別データとして使用した。

３．４．その他データ

日本禁煙学会が2007年8月25日に発表した都道府県別男女別喫煙率データも使用する[10)]。気象データは、理科年表に基づいているが[11)]、気象台や測候所のある都市の気象データをもって、都道府県別のデータとしている。他に社会データも含めており[12) 13)]、標準化死亡比19種類、それ以外222種類、合わせて241種類の都道府県別データを使用することにした。

４．VBAによる重回帰分析ツールの適用

死因データは19種類あるがそれらを全て扱うのは時間その他の制約上困難であるので、ここでは例として脳血管疾患標準化死亡比(男性）だけを重回帰分析することにする。

４．１．脳血管疾患（男性）の標準化死亡比

分析を行う前に、都道府県別の脳血管疾患（男性）の標準化死亡比の棒グラフを図１に示す。標準化死亡比は指数で表示され、100が標準である。沖縄県が78.1で全国最低で、岩手県が136.2で全国最高であり沖縄県の1.74倍である。

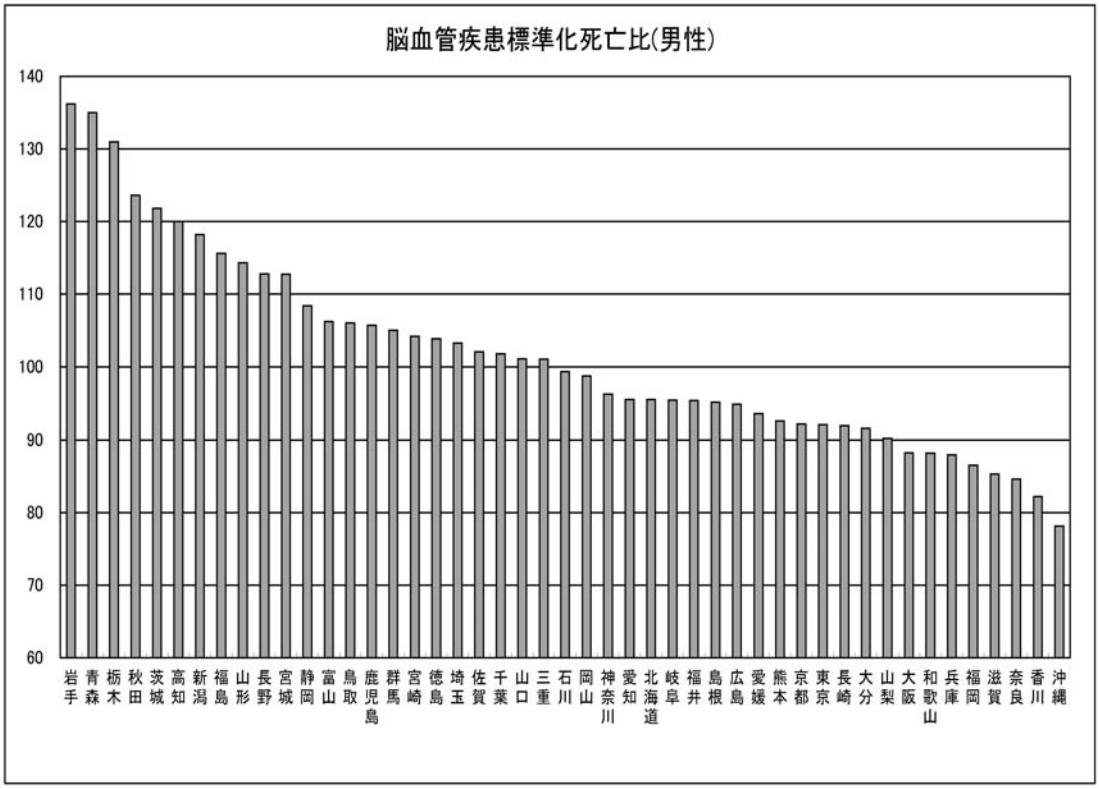

図１　脳血管疾患標準化死亡比（男性）

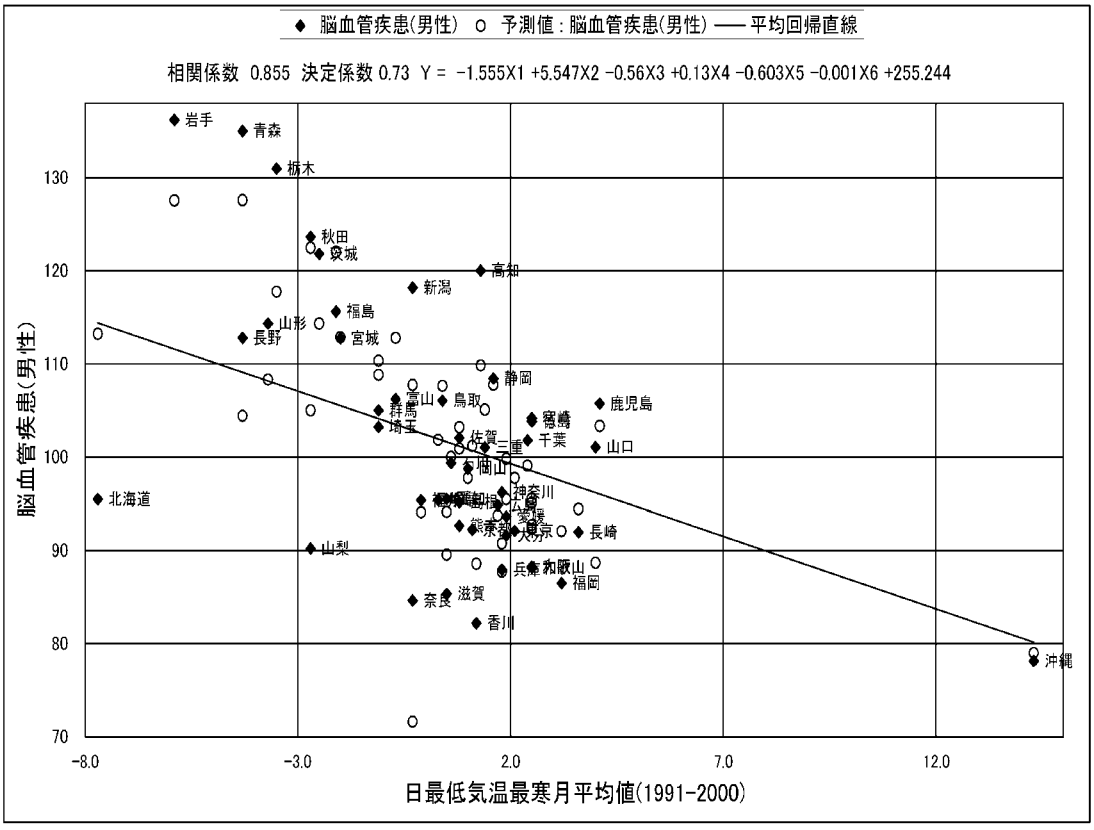

図２　脳血管疾患標準化死亡比（男性）と日最低気温最寒月平均値（1991～2000年平均値）

４．２．単回帰分析による第１説明変数選択

単回帰分析は、説明変数は、「日最低気温最寒月平均値（1991－2000)」を採用した。一日の最低気温の月平均値で年間の最低月の値、一月又は二月の日最低気温の平均値である。決定係数の高いものは、「肉類(g／日)(男性)」があったが、寒冷な気候と脳血管疾患のイベントとの因果関係は明白と考え、まずこれを説明変数として採用する。脳血管疾患死亡率と最寒月の最低気温の月平均値との相関は、既に指摘されているところであり、室外の低温環境への暴露が脳血管疾患発症の原因となる。

図２に、縦軸に脳血管疾患標準化死亡比(男性)、横軸にここで採用した説明変数をとった散布図を示す。図２は、後述の最終的に得られた6個の説明変数による重回帰分析の結果を表しており、図中の黒菱形が各都道府県の実測値、白丸が予測値である。図中の直線は、重回帰分析でも説明変数の効果が分かりやすいように付加した直線で、後の「5.3 平均回帰直線」で説明する。以後、図７まで、同様の構成になっている。

４．３．重回帰分析による第２説明変数選択

説明変数２個の重回帰を行い、第２説明変数として「食塩(g／日)(男性)」を採用した。第１説明変数は、「日最低気温最寒月平均値（1991－2000)」である。

決定係数のより高いものは順に、「肉類(g／日)(男性)」、「前立腺がん（男性）」、「エネルギー(kcal／日)(男性)」、「炭水化物(g／日)(男性)」、「脂質(g／日)(男性)」、「その他の野菜類(g／日)(全体)」、「肝がん(男性)」、「大腸がん(男性)」、「うち動物性脂質(g／日)(男性)」、「豆類(g／日)(男性）」があったが、食塩摂取は高血圧ひいては脳血管疾患の原因とみなせるので、敢えて食塩を採用した。図３に、追加した第２説明変数を横軸にとった散布図を示す。図中の黒菱形は各都道府県の実測値、白丸は最終的に得られた6個の説明変数による重回帰分析による予測値であることは、前と同じである。

４．４．重回帰分析による第３説明変数選択

説明変数３個の重回帰により、第３説明変数として「肉類（g／日)(男性)」を採用した。第1説明変数は、「日最低気温最寒月平均値（1991－2000)」、第２説明変数は「食塩(g／日)(男性)」である。

決定係数のより高いものは順に、「エネルギー(kcal／日)(男性)」があったが、「肉類(g／日)(男性)」との決定係数の違いは極わずかであったこと、トータルカロリーより肉と限定した方が栄養成分のどれがというイメージがつかみ易いことから採用した。図４に、追加した第３説明変数を横軸とする散布図を示す。図の見方は前と同じである。

脳血管疾患は、年齢調整済み死亡率が1965年から1995年までに１／４ないし１／３程度に激減し、これには食事の洋風化が寄与していると考えられていることから、説明変数として「肉類(g／日)(男性)」を採用することは自然であると思われる。

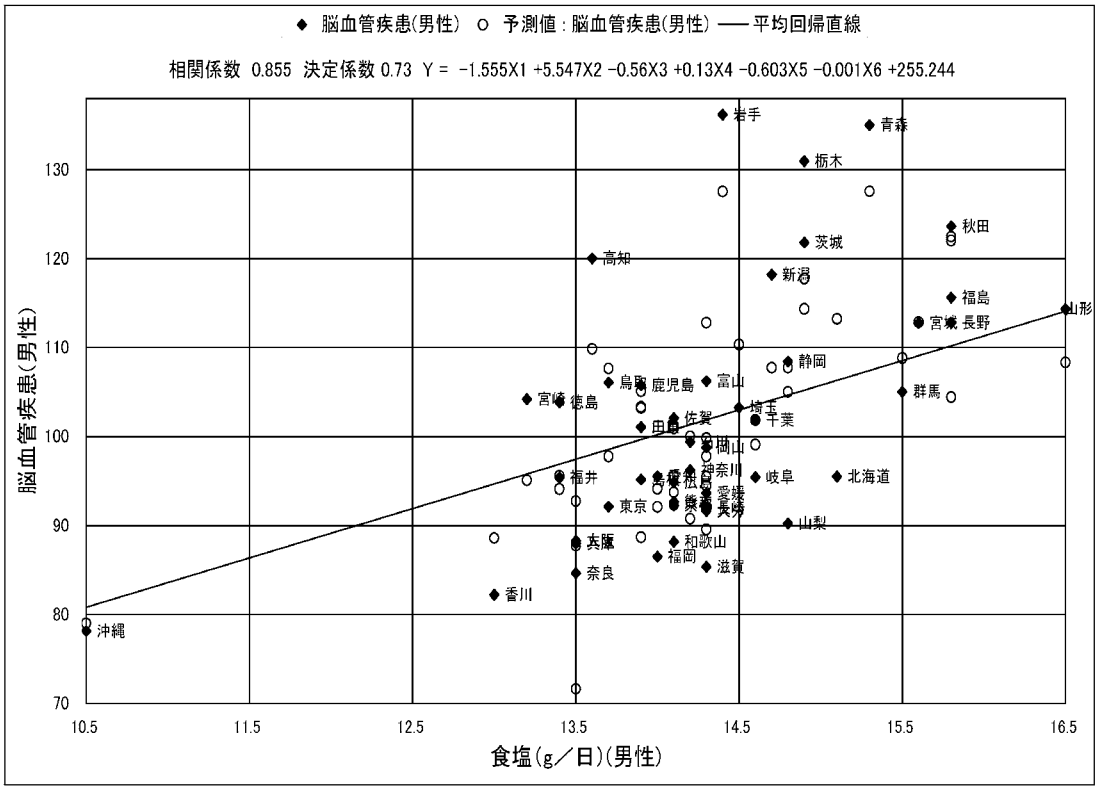

図３　脳血管疾患標準化死亡比（男性）と食塩（g／日）（男性）の散布図

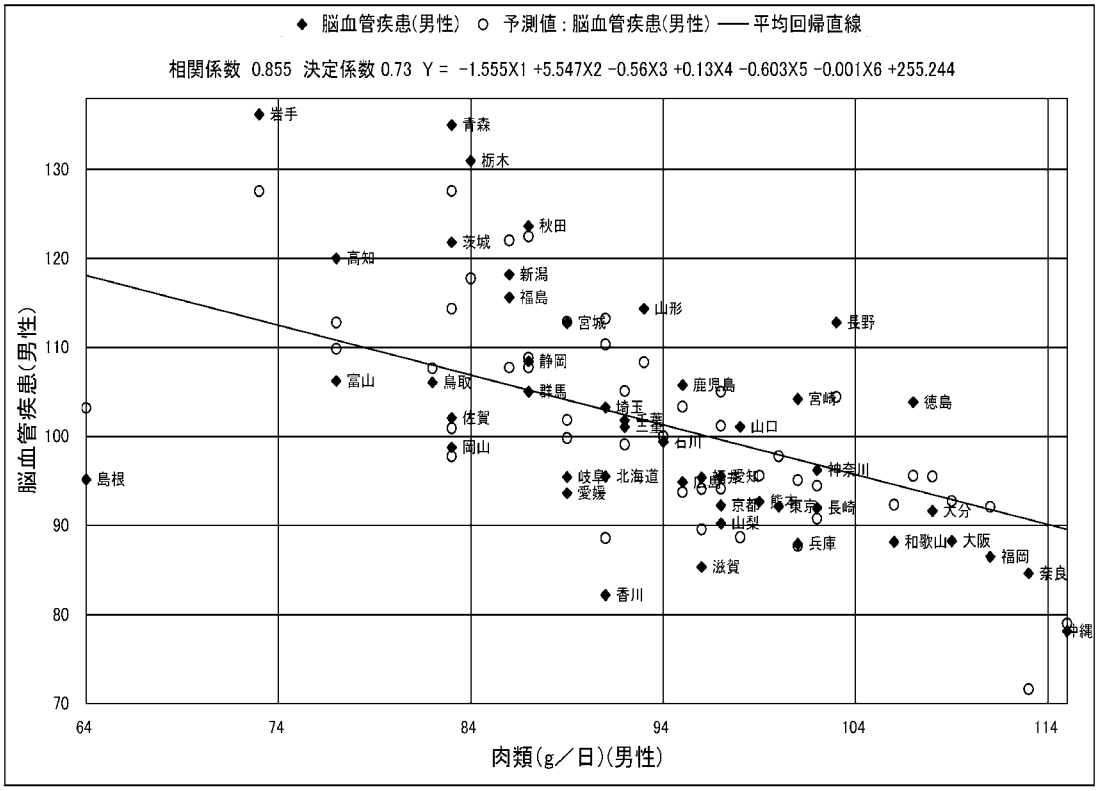

図４　脳血管疾患標準化死亡比（男性）と肉類（g／日）（男性）の散布図

４．５．重回帰分析による第４説明変数選択

説明変数4個の重回帰を行い、第４説明変数として「糖尿病標準化死亡比（男性）」を採用した。第1説明変数は、「日最低気温最寒月平均値（1991－2000）」、第２説明変数は「食塩（g／日）（男性）」、第３説明変数は「肉類（g／日）（男性）」である。

糖尿病は動脈硬化や細小血管の障害原因となり、有症状の脳梗塞や脳塞栓、無症候性脳梗塞等を引き起こすと言われ、脳血管疾患との因果関係があると考えられる。決定係数のより高いものは順に、「炭水化物（g／日）（男性）」、「大腸癌標準化死亡比（男性）」等があったが、糖尿病の方が脳血管疾患（男性）との関係がより確実であると思われ、これを説明変数に採用した。図５に、追加した第４説明変数を横軸とする散布図を示す。

４．６．重回帰分析による第５説明変数選択

説明変数５個の重回帰により、第５説明変数として「炭水化物（g／日）（男性）」を採用した。第１説明変数は、「日最低気温最寒月平均値（1991－2000）」、第２説明変数は「食塩（g／日）（男性）」、第３説明変数は「肉類（g／日）（男性）」、第４説明変数は「糖尿病標準化死亡比（男性）」である。図６に、追加した第５説明変数を横軸とする散布図を示す。

この時点では、「炭水化物（g／日）（男性）」の決定係数は、「エネルギー（kcal／日）（男性）」より0.05高いなど最も高く、これを採用した。説明変数選択では生態学的誤謬の危険が常に付きまとうが、炭水化物摂取については特に因果関係が見出しにくい。

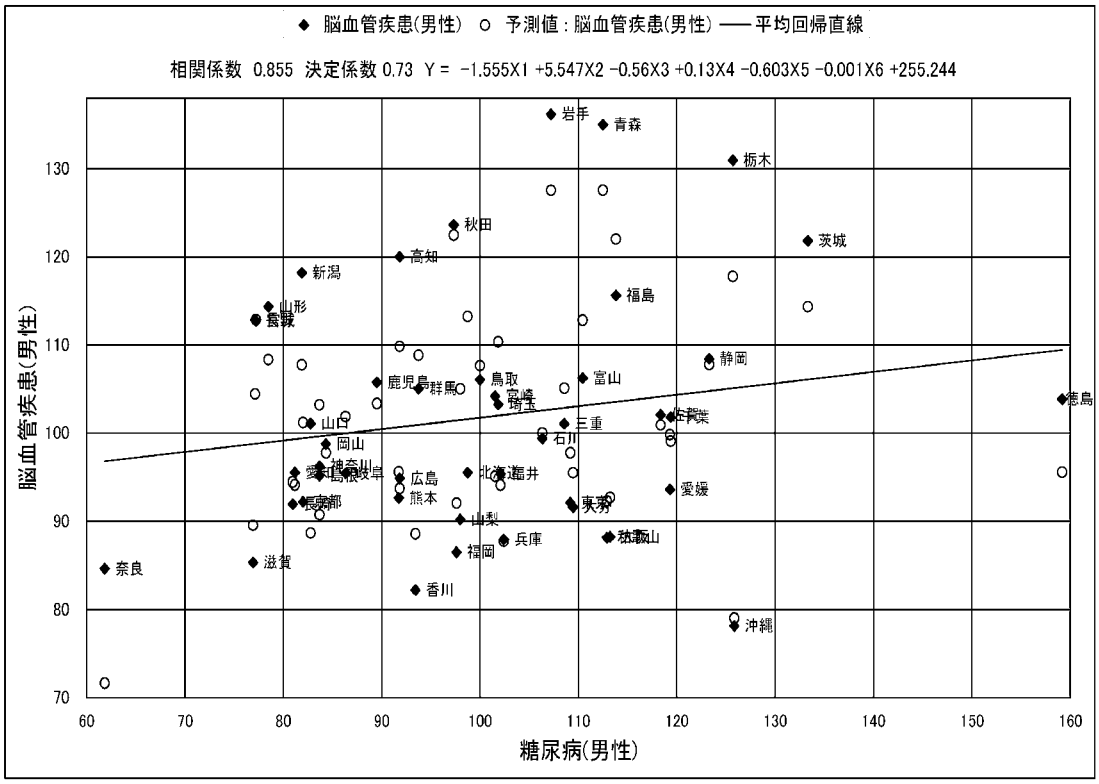

図５　脳血管疾患標準化死亡比（男性）と糖尿病標準化死亡比（男性）の散布図

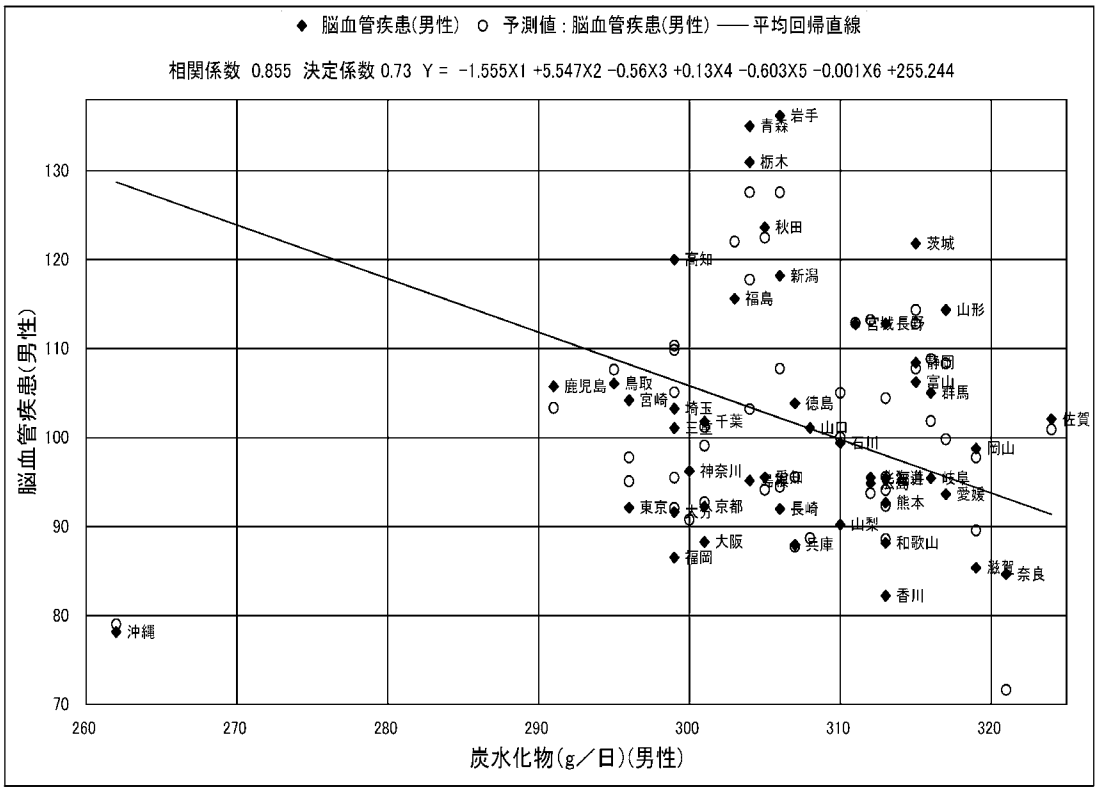

図６　脳血管疾患標準化死亡比（男性）と炭水化物（g／日）（男性）の散布図

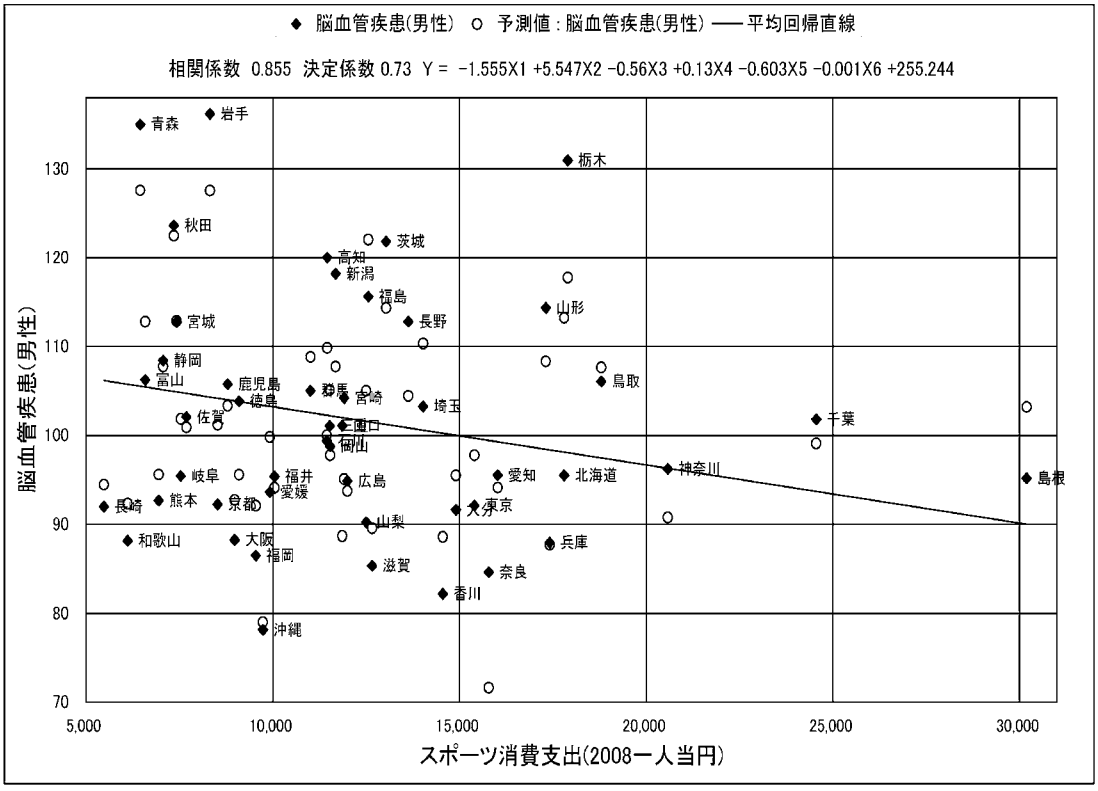

図７　脳血管疾患標準化死亡比（男性）とスポーツ消費支出（一人当り円）（全体）の散布図

４．７．重回帰分析による第６説明変数選択

説明変数６個の重回帰分析により、第６説明変数として「スポーツ消費支出（2008一人当たり円）（全体）」を採用した。第１説明変数は、「日最低気温最寒月平均値（1991－2000）」、第２説明変数は「食塩（g／日）（男性）」、第３説明変数は「肉類（g／日）（男性）」、第４説明変数は「糖尿病標準化死亡比（男性）」、第５説明変数は「炭水化物（g／日）（男性）」と固定した上で、説明変数の候補を探し、決定係数の高い候補として採用した。

図７に、追加した第６説明変数を横軸とする散布図を示す。第６説明変数を導入した後の重回帰式の偏回帰係数のｐ値は、より説明変数の少ない場合より、概して向上している。

４．８．重回帰結果

重回帰全般については、重回帰式の重決定係数は0.73、重相関係数は0.855、重回帰式は、Ｙを予測値、X_1～X_6を第１～第６説明変数として、以下のものであった。

$$Y = -1.55X_1 + 5.55X_2 - 0.560X_3 + 0.130X_4 - 0.603X_5 - 0.000656X_6 + 255.2$$

図２から図７に、各説明変数を横軸として重回帰の結果を既に図示してある。重回帰式のｐ値は、5.31×10^{-10}である。偏回帰係数のｐ値を含め、表1にまとめて示す。

表1　重回帰分析の回帰式および偏回帰係数のｐ値

重回帰式	切　　片	第１説明変数	第２説明変数
5.31×10^{-10}	2.21×10^{-5}	0.0144	0.00765
第３説明変数	第４説明変数	第５説明変数	第６説明変数
4.49×10^{-5}	0.0527	0.000106	0.00916

５．標準化死亡比に対するリスク分析方法、及び平均回帰直線の提案

５．１．ＳＭＲリスク差

標準化死亡比（ＳＭＲ）が被説明変数であるので、発症率やそのロジットを使用しておらず、リスク比を求めることはできない。ここでは、それに代わる代替指標を考えることにする。

ある説明変数、例えばｋ番目の説明変数からのリスクに相当する指標として、都道府県を2群に分け、ｋ番目の説明変数に関する観測値の高い23都道府県（高位群）の平均値（高位平均値）と、低い観測値の23都道府県（低位群）のそれ（低位平均値）を求め、高位平均値に対する標準化死亡比の予測値（SMR高位群予測値）と低位平均値に対するそれ（SMR低位群予測値）の差を求め、これをもって高位群と低位群の２群間でのリスクの違いを示す指標とし、ここでは「ＳＭＲリスク差」と呼ぶこととする。

予測値を求める際、ｋ番目の説明変数は高位又は低位平均値を使うが、ｋ番目以外の

説明変数は、観測値平均値（全ての都道府県の平均値）を使用する。これはｋ番目以外の説明変数の意味するリスクに関してリスク中立の条件を設定したことになる。ｋ番目の説明変数だけは、高位群と低位群で異なる値を使う。これにより、被説明変数への寄与は、ｋ番目の説明変数の意味するリスクのみが取り込まれることになり、高位群と低位群の被説明変数への影響の違いを知ることができる。

以上の手順を箇条書きにすると、以下になる。

① 47都道府県を、ある説明変数ｋ番目の説明変数について、観測値の小さい低位群の23都道府県と観測値の大きい高位群の23都道府県の２群に分ける。

② ｋ番目以外の説明変数は、全都道府県の平均値を用いる。

③ ｋ番目の説明変数は、低位群又は高位群の都道府県の平均値を用いる。

④ 重回帰式には、ｋ番目以外の説明変数は②の平均値を、ｋ番目の説明変数は③の低位群又は高位群の平均値を代入して、予測値を求める。

⑤ ④で求めた高位群と低位群の予測値の差を、２群間のリスク差とみなす。

⑥ ⑤のリスク差（ＳＭＲリスク差）に対して、95%信頼区間を求める。

５．２．ＳＭＲリスク差の95%信頼区間

「ＳＭＲリスク差」の95%信頼区間を合わせて求める。「ＳＭＲリスク差」が、高位群平均値と低位群平均値の差と、ｋ番目の説明変数の偏回帰係数の積である。これより、ＳＭＲリスク差の95%信頼区間は、ｋ番目の説明変数の偏回帰係数の95%信頼区間と、高位群平均値と低位群平均値の差との積となる。

５．３．平均回帰直線

重回帰分析の場合は、単回帰分析と異なって説明変数との直線関係のような視覚的にわかりやすい関係はない。２個以上の説明変数による被説明変数の予測値と観測値のグラフ上のプロットは、関係を見出しにくいものとなる場合がよくある。

そこで、特定の説明変数、例えばｋ番目の説明変数の影響をわかりやすく示すため、重回帰式において、ｋ番目以外の説明変数は全て各説明変数の平均値を使い（即ちｋ番目以外の説明変数の意味するリスクについてリスク中立の条件を課し）、ｋ番目の説明変数は独立変数として、被説明変数の予測値との関係を直線で示すことにする。これを「平均回帰直線」と呼ぶこととし、重回帰分析の結果のグラフに付加する。この直線により、ｋ番目の説明変数の効果を視覚的に分かり易く捉える事が可能となる。

５．４．ＳＭＲリスク差・95%信頼区間・平均回帰直線等の解析的表現

５．４．１．偏回帰係数の期待値

ｙを被説明変数の観測値、x_iをｉ番目の説明変数の観測値、βを偏回帰係数、εを誤差、ｎを都道府県の数、ｐを説明変数の数とする。重回帰分析の解説書[14] [15]もあるが、以下はこれらとは異なった説明の仕方をしている。

$y = (y_1, y_2, \cdots, y_n)^t$ 、 $x_i = (x_{1i}, x_{2i}, \cdots, x_{ni})^t$ 、 $\beta = (\beta_0, \beta_1, \cdots, \beta_p)^t$ 、

$\varepsilon = (\varepsilon_1, \epsilon_2, \cdots, \epsilon_n)^t$

$$y = \sum_{i=1}^{p} \beta_i x_i + \beta_0 + \varepsilon$$

Ｙを予測値とし$Y = (Y_1, Y_2, \cdots, Y_n)^t$で表し、$\hat{\beta}$を偏回帰係数の期待値とし、$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p)^t$で表すものとする。

$$Y = \sum_{i=1}^{p} \hat{\beta} x_i + \hat{\beta}_0 \quad \text{ここに} \ \hat{\beta} = (X^t X)^{-1} X^t y$$

$\hat{\beta}$は最小二乗法から求めた偏回帰係数の期待値である。

ただし、$X = [1_n, x_1, \cdots, x_p]$、Xは行列でn行（p+1）列、$1_n = (1, \cdots, 1)^t$（ｎ次元ベクトル）とする。$X$と$X^t X$は、行または列を第０行（第０列）からカウントする。

５．４．２．ＳＭＲリスク差

次にｋ番目の説明変数について、値の大きい23都道府県即ち高位群の平均値を$\bar{x}_{UP}$、高位群予測値をY_{UP}とし、値の小さい23都道府県即ち低位群の平均値を$\bar{x}_{LOW}$、低位群予測値をY_{LOW}として、２群の予測値の差を求める。被説明変数の平均値を$\bar{y}$、ｉ番目の説明変数の平均値を$\bar{x}_i$とすると、

$\bar{y} = \frac{1}{n}\sum_{\mu=1}^{n} y_\mu$、　　$\bar{x}_i = \frac{1}{n}\sum_{\mu=1}^{n} x_{\mu i}$であり、

$$Y_{LOW} = \sum_{i \neq k}^{p} \hat{\beta}_i \bar{x}_i + \hat{\beta}_k \bar{x}_{LOW} + \hat{\beta}_0 = \sum_{i=1}^{p} \hat{\beta}_i \bar{x}_i + \hat{\beta}_0 + \hat{\beta}_k (\bar{x}_{LOW} - \bar{x}_k) = \bar{y} + \hat{\beta}_k (\bar{x}_{LOW} - \bar{x}_k)$$

$$Y_{UP} = \sum_{i \neq k}^{p} \hat{\beta}_i \bar{x}_i + \hat{\beta}_k \bar{x}_{UP} + \hat{\beta}_0 = \sum_{i=1}^{p} \hat{\beta}_i \bar{x}_i + \hat{\beta}_0 + \hat{\beta}_k (\bar{x}_{UP} - \bar{x}_k) = \bar{y} + \hat{\beta}_k (\bar{x}_{UP} - \bar{x}_k)$$

ここで、$\bar{y} = \sum_{i=1}^{p} \hat{\beta}_i \bar{x}_i + \hat{\beta}_0$という関係を使った。

$$E(Y_{UP} - Y_{LOW}) = E(\beta_k (\bar{x}_{UP} - \bar{x}_{LOW})) = (\bar{x}_{UP} - \bar{x}_{LOW}) \hat{\beta}_k$$

ここに、$E(Y_{UP} - Y_{LOW})$は「ＳＭＲリスク差」である。

５．４．３．ＳＭＲリスク差の95%信頼区間

ＳＭＲリスク差の分散については、以下となる。

$$V(Y_{UP} - Y_{LOW}) = E\left(((Y_{UP} - Y_{LOW}) - E(Y_{UP} - Y_{LOW}))^2\right) = (\bar{x}_{UP} - \bar{x}_{LOW})^2 E((\beta_k - \hat{\beta}_k)^2)$$

$E\left((\beta - \hat{\beta})(\beta - \hat{\beta})^t\right) = \sigma^2 (X^t X)^{-1}$より、$E((\beta_k - \hat{\beta}_k)^2) = \sigma^2 ((X^t X)^{-1})_{kk}$

ただし、σを誤差の標準偏差とし、$E(\varepsilon \varepsilon^t) = \sigma^2 E_n$（$E_n$はｎ次単位行列）とした。

$T = \dfrac{(\beta_k - \hat{\beta}_k)}{\sigma_e \sqrt{((X^t X)^{-1})_{kk}}}$とおくと、Tは自由度ｎ－ｐ－１のｔ分布に従う。

実際に、Tの満たす確率密度関数g(T)を求めることにする。σ_eを下式で示される自由度ｎ－ｐ－１の誤差の標本標準偏差とし、ωを定義する。これらを使いTを書き換える。

$$\sigma_e^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - Y_i)^2 \text{、} \omega = {\sigma_e}^2(n-p-1)/\sigma^2$$

ji

$$T = \frac{(\beta_k - \hat{\beta}_k)/\sigma\sqrt{((X^tX)^{-1})_{kk}}}{\sigma_e/\sigma} = \frac{(\beta_k - \hat{\beta}_k)/\sigma\sqrt{((X^tX)^{-1})_{kk}}}{\sqrt{\omega/(n-p-1)}}$$

ωは自由度ｎ－ｐ－１のカイ二乗分布に従うことから、正規分布に関する積分等で置き換え、g(T)が以下の様に求まる。δはデルタ関数、Γはガンマ関数である。

$$\mathrm{g(T)}=\int_0^{\infty} d\omega\left[\int_{-\infty}^{\infty}\delta(T-\frac{(\beta_k-\hat{\beta}_k)/\sigma\sqrt{((X^tX)^{-1})_{kk}}}{\sqrt{\omega/(n-p-1)}})\frac{1}{\sqrt{2\pi((X^tX)^{-1})_{kk}}\sigma}\exp(-\frac{(\beta_k-\hat{\beta}_k)^2}{2\sigma^2((X^tX)^{-1})_{kk}})\mathrm{d}\beta_k\right.$$

$$\left.\times\int_{-\infty}^{\infty}\delta\left(\omega-\sum_{\mu=1}^{n-p-1}{z_\mu}^2\right)\prod_{\mu=1}^{n-p-1}\frac{1}{\sqrt{2\pi}}\exp(-\frac{{z_\mu}^2}{2})\,\mathrm{d}z_\mu\right]$$

$$=\frac{1}{\sqrt{\pi(n-p-1)}}\frac{\Gamma(\frac{n-p}{2})}{\Gamma(\frac{n-p-1}{2})}\left(1+\frac{T^2}{n-p-1}\right)^{-\frac{n-p}{2}}$$

g(T)は、自由度ｎ－ｐ－１のｔ分布の確率密度関数となっている。

ｔ（･,･）をｔ分布の累積確率分布関数、α=0.05、ｎ－ｐ－１を誤差の重回帰を除く自由度として、$\hat{\beta}_k$の95%信頼区間は、ｔ分布より$[\hat{\beta}_k \pm t(1-\alpha/2, n-p-1)\,\sigma_e\sqrt{((X^tX)^{-1})_{kk}}]$となる。このことから、$E(Y_{UP}-Y_{LOW})$の95%信頼区間は、以下のようになる。

$$\left[(\bar{x}_{UP}-\bar{x}_{LOW})\hat{\beta}_k \pm (\bar{x}_{UP}-\bar{x}_{LOW})t\left(1-\frac{\alpha}{2}, n-p-1\right)\sigma_e\sqrt{((X^tX)^{-1})_{kk}}\ \right];\ 1\le k\le p$$

５．４．４．平均回帰直線

平均回帰直線は、ｋ番目を除く説明変数は観測値の平均値を、ｋ番目の説明変数の値をｘとし、予測値を$Y_{REGRESS}$として、ｋ番目の説明変数の影響を表現する平均回帰直線は以下の式になる。

$$Y_{REGRESS} = \sum_{i\neq k}^{p}\hat{\beta}_i\bar{x}_i + \hat{\beta}_k x \quad + \hat{\beta}_0 = \sum_{i=1}^{p}\hat{\beta}_i\bar{x}_i + \hat{\beta}_0 + \hat{\beta}_k(x-\bar{x}\ ) = \bar{y} + \hat{\beta}_k(x-\bar{x}_k)$$

平均回帰直線の予測値の95%信頼区間は、ｋ番目の説明変数以外は観測値の平均値を使用していることから、以下となる。信頼区間はｋ番目の説明変数の値ｘにのみ依存する。

$$\left[Y_{REGRESS} \pm t\left(1-\frac{\alpha}{2}, n-p-1\right)\sigma_e\sqrt{\frac{1}{n}+(x-\bar{x}_k)^2((X^tX)^{-1})_{kk}}\ \right];\ 1\le k\le p$$

５．４．５．行列（X^tX）$^{-1}$の次数を１つ下げた表現

さらに、$A = X^t\mathrm{X}$とおき、$A = \begin{bmatrix} A_{00} & A_{01} \\ A_{10} & A_{11} \end{bmatrix}$とすると、

A_{00}= n 、 $A_{10} = {A_{01}}^t = (n\bar{x}_{1,\cdots,}n\bar{x}_{p,})^t$、 $(A_{11})_{ij} = (\sum_{\mu=1}^{n} x_{\mu i}x_{uj})_{ij}$ $1 \le i,j \le p$である。逆行列を

$A^{-1} = \begin{bmatrix} (A^{-1})_{00} & (A^{-1})_{01} \\ (A^{-1})_{10} & (A^{-1})_{11} \end{bmatrix}$として、さらにDをｐ次実対称行列として、

$$D = \begin{bmatrix} \sum_{\mu=1}^{n}(x_{\mu 1}-\bar{x}_1)^2 & \cdots & \sum_{\mu=1}^{n}(x_{\mu 1}-\bar{x}_1)(x_{\mu p}-\bar{x}_p) \\ \vdots & \vdots & \vdots \\ \sum_{\mu=1}^{n}(x_{\mu p}-\bar{x}_p)(x_{\mu 1}-\bar{x}_1) & \cdots & \sum_{\mu=1}^{n}(x_{\mu p}-\bar{x}_p)^2 \end{bmatrix}$$ とする。Dを使って、

$$(A^{-1})_{00} = \frac{1}{n} + \sum_{i,j=1}^{p} \bar{x}_i\bar{x}_j(D^{-1})_{ij} \ ; \ (A^{-1})_{11} = D^{-1} \ ;$$

$$(A^{-1})_{10} = {(A^{-1})_{01}}^t \ ; \qquad ((A^{-1})_{01})_j = -\sum_{i=1}^{p} \bar{x}_i \ (D^{-1})_{ij} \qquad 1 \le j \le p$$

これから$((X^tX)^{-1})_{kk}$は次のように表現される。

$((X^tX)^{-1})_{kk} = ((D)^{-1})_{kk}$　$1 \le k \le p$

５．５．重回帰分析結果へのリスク分析の適用

脳血管疾患標準化死亡比（男性）に対する、重回帰分析結果のＳＭＲリスク差・95%信頼区間などをまとめて、表２に示す。ＳＭＲリスク差は、標準化死亡比（ＳＭＲ）の高位群予測値と低位群予測値の差である。標準化死亡比は100を標準とする指数であり、ＳＭＲリスク差の値は%に相当する。

表２　脳血管疾患標準化死亡比（男性）のＳＭＲリスク差と95%信頼区間

変数 / リスク関係指標	被説明変数	説明変数１	説明変数２	説明変数３	説明変数４	説明変数５	説明変数６
	脳血管疾患標準化死亡比（男性）	日最低気温最寒月平均値（1991-2000）	食塩（ｇ／日）（男性）	肉類（ｇ／日）（男性）	糖尿病標準化死亡比（男性）	炭水化物（ｇ／日）（男性）	スポーツ消費支出（2008一人当円）
中央値	99.4	0.8	14.2	93.0	98.7	307.0	11537
高位群平均値	112.6	2.7	14.9	101.8	114.6	314.3	15962
低位群平均値	91.0	-1.8	13.6	84.6	85.4	299.2	8571
SMR高位群予測値		98.3	105.4	96.9	103.6	97.2	99.3
SMR低位群予測値		105.2	98.1	106.6	99.8	106.3	104.2
ＳＭＲリスク差		-7.0	7.3	-9.6	3.8	-9.1	-4.8
SMR95%信頼区間下限		-12.5	2.0	-13.9	0.0	-13.4	-8.4
SMR95%信頼区間上限		-1.5	12.5	-5.4	7.7	-4.8	-1.3
偏回帰係数ｐ値再掲		0.0144	0.00765	4×10^{-5}	0.0526	1×10^{-4}	0.009

表2にあるように、高位群と低位群を分析して、以下の結果となった。[ ]内は、95%信頼区間である。標準化死亡比は100を標準とする指数であり、以下では%表示をする。

① 日最低気温最寒月平均値（1991－2000）は、高位群（気温が高い方）が7.0%（[1.5－12.5]%）死亡比が低い。

② 食塩（g／日）（男性）摂取では、高位群が7.3%（[2.0－12.5]%）死亡比が高い。

③ 肉類（g／日）（男性）摂取では、高位群が9.6%（[5.4－13.9]%）死亡比が低い。

④ 糖尿病標準化死亡比（男性）では、高位群が3.8%（[0.0－7.7]%）死亡比が高い。

⑤ 炭水化物（g／日）（男性）摂取では、高位群が9.1%（[4.8－13.4]%）死亡比が低い。

⑥ スポーツ消費支出（2008一人当円）では、高位群が4.8%（[1.3－8.4]%）死亡比が低い。

このうち、脳血管疾患標準化死亡比と食塩摂取との関係であるが、食塩摂取が高位の群が脳血管疾患の死亡比が高くなるのはわかるが、高位摂取群14.9gと低位摂取群13.6gとで、これだけのリスク差がでるものか疑問が残る。

また、既に記したように、炭水化物の摂取と脳血管疾患による死亡との間に、筆者には明瞭な関係を見出すことが難しい。

以上の点はあるが、その他の結果は、絶対的数値はともかく、傾向に関しては容易に推察できるようなものであったといえよう。即ち、全般として、リスク比などの既存のリスク分析の代替となりうるような、受け入れやすいリスクの表現方法になっていると、筆者は考えている。

5．6．重回帰分析結果へ「平均回帰直線」の適用

「5.3　平均回帰直線」で記したように、重回帰分析の場合は、説明変数による被説明変数の予測値への効果が見出しにくくなりがちであり、特定の説明変数以外は全て各説明変数の平均値を使い、特定の説明変数は独立変数として、被説明変数の予測値との関係を直線で示すものである。これは、既に重回帰分析結果の図2から図7の散布図中に直線で示してあり、各図の横軸にとった説明変数の効果を、わかり易く視覚化できたものと筆者は考えている。

6．結論

VBAによる重回帰分析ツールを健康データに適用した。具体的には、都道府県別脳血管疾患標準化死亡比（男性）を、都道府県別栄養摂取データを中心に約240の都道府県別データを使って重回帰分析を行った。6個の説明変数を使用した重回帰分析例を示した。その結果に対して、説明変数毎に高位と低位の都道府県2群に分けて標準化死亡比の差を求めた。また95%信頼区間も求めた。

これにより、VBAで開発した重回帰分析ツールを、標準化死亡比と栄養摂取データ等による分析に実際に適用可能であることを示した。また、2群間でリスク差の評価、及び信頼区間の評価に関する試案もVBAで実装し、可能性を提示した。重回帰分析で特定の説明変数の影響を視覚的に明快に示す「平均回帰直線」とここで呼ぶものを提案し、

実際に使用し、望ましい結果を得たと考えている。既に記してきたとおりである。

7．今後の課題

まず必要なことは、重回帰分析ツールの適用例が本稿だけであるので、もっと適用例を増やしていき、重回帰分析ツールの有用性を多くの事例をもとに検討していくことが望まれる。適用例が増えれば、ツールの改良等の必要性も具体的に把握できてくるものと思われる。ともかく、適用例を増やしてみるのが先決であろう。

なお、摂取した食物・栄養と標準化死亡比との関係が見られるとき、因果関係の可能性がありそうか否かの判断が難しい。説明変数の候補となる都道府県別データが240種程度あると、たまたま標準化死亡比と関連があるように見えるデータが、意外に多く存在する。地域相関研究は因果関係の可能性を示唆するにすぎないが、重回帰の決定係数が偶然大きく出たデータの因果関係の可能性を手早く識別する方法の開発が必要である。

さらに、本稿では都道府県別データは、47都道府県分を47個のデータとみなしている。要するに、例えば東京都のデータを人口に応じて、重みを付加して取り扱うことはしていない。データの重みの調整ができれば特定の都道府県別を指定しておいて、フィッティングを良くすることができるかも知れない。

参考文献

1）堀田裕史：「VBAによるユーザーの判断を重視した重回帰分析ツールの開発」、富山短期大学紀要、Vol.43（1）、pp.115－130（2008）.

2）高俊珂，梯正之：「都道府県別の平均寿命と社会・経済指標および栄養指標との関連性」、広島大学保健学ジャーナル、Vol. 5、pp.62－69（2006）.

3）日本国際生命科学協会砂糖研究部会編：『栄養疫学　可能性と限界』、日本国際生命科学協会（1998）

4）佐々木敏：『わかりやすいEBNと栄養疫学』、同文書院（2005）

5）Walter Willett著、田中平三監訳：『食事調査のすべて－栄養疫学－第2版』、第一出版（2003）

6）厚生労働省老健局老人保健課：「都道府県別死因の分析結果について」、http://www.mhlw.go.jp/topics/2005/02/tp0228-2/（2009年8月21日現在）

7）厚生労働省老健局老人保健課：「標準化死亡比データファイル」、http://www.mhlw.go.jp/topics/2005/02/tp0228-2/xls/gf1.xls（2009年8月21日現在）

8）中村美詠子他：「国民栄養調査を活用した都道府県別栄養関連指標の検討」、http://www2.hama-med.ac.jp/w1a/health/jouho/eiyoushihyou/h14nnss.pdf 、http://www.nih.go.jp/eiken/yousan/eiyochosa/（2009年8月25日現在）

9）総務省統計局：「家計調査　（品目分類）第11表　都市階級・地方・都道府県庁所在市別１世帯当たり年間の品目別支出金額（総世帯）　平成20年」

http://www.e-stat.go.jp/SG1/estat/NewList.do?tid=000001042133
（2009年8月25日現在）
10）ＮＰＯ法人日本禁煙学会：「都道府県別男女別喫煙率」、
http://www.nosmoke55.jp/data/0708todoufuken.pdf（2009年8月25日現在）
11）東京天文台編：『理科年表 2005』、丸善株式会社（2004）
12）（財）矢野恒太郎記念館編：『県勢ＣＤ－ＲＯＭ　2004』、（財）矢野恒太郎記念館（2003）
13）経済企画庁編：『平成11年版　新国民生活指標』、大蔵省印刷局（1999）
14）柳井晴夫、高木廣文：『多変量解析ハンドブック』、現代数学社（1986）
15）丹後俊郎：『新版　医学への統計学』、朝倉書店（1993）

（平成21年10月30日受付、平成21年11月９日受理）