回帰分析の研究
―検定を中心に―

2010SE217 高田貴文
指導教員：木村美善

1　はじめに

本研究の目的は回帰分析における仮説検定の理解をすることである．理論的内容の学習については [3] を参考とし，[1] と [2] は補助的に利用した．さらに理論の内容をより深く理解するために [4] のデータを用いて検定を行った．そして，線形制約のある基本的な検定モデルと特殊な形のモデルに対して実際のデータを適用することによって，検定の有用性について考察を行った．

1.1　回帰モデル

目的変数を y，説明変数を $x_j \quad (j = 1, ..., p)$ とし，

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, \cdots, n$$

を考える．ただし，$\beta_0, \beta_1, \cdots, \beta_p$ は回帰係数を表し，ε_i は誤差項である．以下，これを行列，ベクトル表記する．目的変数からなる $n \times 1$ 行列を $\boldsymbol{y}$, 定数項と説明変数からなる $n \times (p+1)$ 行列を X，回帰係数の $(p+1) \times 1$ ベクトルを $\boldsymbol{\beta}$ とすると

$$y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

と表せる．ここで，

$$E(\boldsymbol{\varepsilon}) = 0, \quad V(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I} \tag{2}$$

を仮定する．(2) に対して，次の $\boldsymbol{\beta}$ に関する線形制約条件

$$H'\boldsymbol{\beta} = \boldsymbol{\xi}_0 \tag{3}$$

が満たされる場合を考える．H は既知の定数を要素とする $p \times r$ 行列であり，r 個の列は 1 次独立，すなわち $\mathrm{rank}H = r(< p)$ と仮定する．

一般的な回帰モデルの特殊な場合として次の 3 種類の検定問題を考える．

(a) 回帰式の同等性の検定

$$y_{(1)} = X_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\varepsilon}_{(1)}, \boldsymbol{\varepsilon}_{(1)} \sim N(0, \sigma^2 I) \tag{4}$$
$$y_{(2)} = X_{(2)}\boldsymbol{\beta}_{(2)} + \boldsymbol{\varepsilon}_{(2)}, \boldsymbol{\varepsilon}_{(2)} \sim N(0, \sigma^2 I) \tag{5}$$

・帰無仮説 $H_0 : \boldsymbol{\beta}_{(1)} = \boldsymbol{\beta}_{(2)} (= \boldsymbol{\beta})$ の対立仮説 $H_1 : \boldsymbol{\beta}_{(1)} \neq \boldsymbol{\beta}_{(2)}$ に対する検定統計量

$$W_0 = \frac{RSS_0 - RSS}{RSS} \div \frac{p}{n - 2p} \tag{6}$$

は H_0 のもとで自由度 $(p, n-2p)$ の F 分布に従う．

(b) 回帰式の有意性の検定

一般的な回帰モデル (2) を使用する．

・帰無仮説 $H_0 : \beta_2 = \cdots = \beta_p = 0$ の対立仮説 $H_1 : \beta_2 \neq \cdots \neq \beta_p \neq 0$ に対する検定統計量

$$W_0 = \frac{RSS_0 - RSS}{RSS} \div \frac{p-1}{n-p} = \frac{R^2}{1-R^2} \div \frac{p-1}{n-p} \tag{7}$$

は H_0 のもとで自由度 $(p-1, n-p)$ の F 分布に従う．

(c) 一部の回帰式の同等性の検定

$$\boldsymbol{y}_{(1)} = X_1^{(1)}\boldsymbol{\beta}_1^{(1)} + X_2^{(1)}\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_{(1)} \tag{8}$$
$$\boldsymbol{y}_{(2)} = X_1^{(2)}\boldsymbol{\beta}_1^{(2)} + X_2^{(2)}\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_{(2)} \tag{9}$$

・帰無仮説 $H_0 : \boldsymbol{\beta}_2^{(1)} = \boldsymbol{\beta}_2^{(2)}$ の対立仮説 $H_1 : \boldsymbol{\beta}_2^{(1)} \neq \boldsymbol{\beta}_2^{(2)}$ に対する検定統計量

$$W_0 = \frac{RSS_0 - RSS}{RSS} \div \frac{r}{n-p-r} \tag{10}$$

は H_0 のもとで自由度 $(r, n-p-r)$ の F 分布に従う．

2　データ分析

2.1　データ

実際に [4] より収集した企業データ (四季報 2013 年度) を使用し，企業情報から平均年収を知るための回帰式を導き出すことを目的とし分析を行う．データは，四季報に載っている東証一部上場で正確な企業データが公表されているものを四季報掲載順に使用する．設立年数 (x_1)，資本金 (x_2)，従業員数 (x_3)，売上高 (x_4)，平均年収 (x_5)，平均年齢 (x_6)，平均勤続年数 (x_7)，純利益 (x_8)，採用人数 (x_9) を取り上げ，目的変数を x_5 とし，残りの 8 個を説明変数とした．サンプルデータ数は 55 個である．

2.2　データ全体の分析と分類

まず，収集した全データに対して回帰式を求める分析を行った．しかし，分析結果を見てみると決定係数は 0.503 と低く，全データから分析を行うことはよくないようである．よい分析が出来ない原因として「企業データの中に明らかな偏りを持つ企業がある」ことが考えられるので，次の基準より企業データの分類を行った．

- 企業の分野によって業界分けを行う
- 企業規模を売上高 (1000〜3000 億円) で制限する

この分類より，自動車部品業界 (データ数 15 個)，IT システム業界 (データ数 15 個) の 2 つのグループとした．

2.3 モデル (a) による分析

業界の差があるかを調べるために回帰式の同等性の検定を行う．
$X_{(1)}$ を自動車部品業界に対応する説明変数のデータ
$X_{(2)}$ を IT システム業界に対応する説明変数のデータ
として検定統計量の値を求める．各 RSS_0 と RSS は

$$RSS_0 = 81036.9,\ RSS = 15636.6 \qquad (11)$$
$$n = n_1 + n_2 = 30,\ p = 9$$

となり，各値を代入すると，$W_0 = 5.57$ が得られる．p 値は 0.0037 となる．よって 2 つの業界の間には差があるという仮説はデータより確認された．

2.4 モデル (b) による分析

分類した各グループに対して分析を行い，得られた回帰式に対して検定を行う．
・自動車部品業界の分析，検定を行う．step による変数選択を行い，変数 x_1, x_3, x_6, x_9 が残った．この結果に対して検定統計量の値を求める．

$$R^2 = 0.7807,\ n = n_1 + n_2 = 15,\ p = 5 \qquad (12)$$

となり，これを代入すると $W_0 = 8.9$ が求められる．p 値は 0.0025 となり，帰無仮説は有意水準 1 % において棄却される．得られた回帰式は有意であるといえる．

・IT システム業界も同様に分析，検定を行う．step による変数選択を行い，変数 x_3, x_6, x_7, x_8, x_9 が残った．この結果に対して検定統計量の値を求める．

$$R^2 = 0.9693,\ n = n_1 + n_2 = 15,\ p = 6 \qquad (13)$$

これを代入すると $W_0 = 56.83$ が求められる．p 値は 1.6×10^{-6} となり，帰無仮説は有意水準 1 % において棄却される．得られた回帰式は有意であるといえる．

2.5 ダミー変数を使った業界間の分析

得られた分析結果を参照しながら企業の業界における差異について考える．第一段階としてダミー変数を用いて業種間に違いがあるかどうかを検証する．各グループに新たにダミー変数を自動車部品業界 $\to 1$，IT システム業界 $\to 0$ として追加し，回帰分析を行う．分析の中でダミー変数の係数値，p 値などから回帰式にどれぐらい影響しているかを見る．p 値が低い場合，ダミー変数の影響が大きい (業界の差が大) ということになる．分析の結果は次の通りである．

- 変数選択前の分析から変数全てを使用した場合でもダミー変数の p 値はかなり低い．
- 得られた回帰式からダミー変数は他の変数よりも大きな影響力を持っていることがわかる．
- 変数選択後も決定係数は 0.87 と高く回帰式の信頼性は高い．

これらのことからダミー変数の影響は大きく，2 つのグループに業界の差がみられることがわかった．

2.6 モデル (c) による分析

ここでの目的はダミー変数に対して同等性の検定を行うことにより，前述の方法とは異なる形で業界間の差を検証することである．使用する変数は 2 つの回帰式で共通して残った従業員数，平均年齢，採用人数とし，これをそれぞれのグループに使用して検定を行う．ここでは採用人数 (x_9) を使用した場合のみを記載する．(c) のモデルに以下の前提を加える．

- $\boldsymbol{\beta}_1^{(1)} = \boldsymbol{\beta}_1^{(2)} (= \boldsymbol{\beta}_1)$
- $X_1^{(1)}$ を自動車業界の説明変数 x_{10}
- $X_1^{(2)}$ を自動車業界のダミー変数
- $X_2^{(1)}$ を IT システム業界の説明変数 x_{10}
- $X_2^{(2)}$ を IT システム業界のダミー変数

これに基づき検定統計量の値を求める．各 RSS_0 と RSS は

$$RSS_0 = 302265.9,\ RSS = 261588.6 \qquad (14)$$
$$n = n_1 + n_2 = 30,\ p = 2,\ r = 1$$

となり，各値を代入すると，$W_0 = 4.2$ となる．p 値は 0.0503 となり，帰無仮説は有意水準 5 %において棄却されるかされないかというところである．また，他の共通の変数，従業員数，平均年齢を使用した場合の検定も行った．検定の結果から両方の変数とも有意水準 1 %で帰無仮説が棄却され，業界間の差があることが分かった．

3 おわりに

研究の実施内容として，当初予定していた通り回帰分析の基本的な理論の学習を進め，仮説検定についても多くの線形制約を持つ場合に対応した理論を理解をすることが出来た．しかし，実際にデータを使用した分析と検定においては予定していたところまで進むことが出来なかった．データの解析と各検定は [1] を参考にしながら行ったが，[4] のデータを使用したグループの分類と得られた回帰式の有意性については確認できた．ただし，同等性検定については 2 変数の場合のみの検定になってしまった．3 個以上の変数を使用した場合の検定を行うことが出来なかったのは残念である．

参考文献

[1] Chatterjee，S．and Price，B．：回帰分析の実際，新曜社，1981(加納悟・佐和隆光，原著 1977)．

[2] Rencher，A．C．and Schaalje，G．B．：Linear Models in Statistics，John Wiley，2008．

[3] 佐和　隆光：回帰分析，朝倉書店，1979．

[4] 就活四季報，東洋経済，2013，