§６．質的データの分析

§6.1. 説明変数が質的データである回帰モデル

回帰モデル　$y_i = \alpha + \beta x_i + \varepsilon_i$　において、説明変数x_iがカテゴリーなどの質的データである場合を考えよう。趣味や旅行消費を保有しているバッグのブランド（グッチ、プラダ、セリーヌ等）によって説明するモデルとか、より現実的には、賃金が学歴（大卒、高卒、中卒）によって異なるモデルなどが、それに相当する。

学歴やブランド、属性などのカテゴリー変数は、数値化される（数値で置きかえられる）が、その数値には特別な意味がないのが普通である。例えば、グッチを１、プラダを２、セリーヌを３と数値化したとしても、この１、２、３はブランドの稀少性を示すものではないし、セリーヌがグッチの３倍もの顕示的効果があることを意味しない。

ところが、このような数値化（数量化）が行なわれた説明変数を用いた回帰モデルを、よく考えもせずに推定すると、結果の解釈を誤りやすい状況に陥る。いま、賃金y_iを学歴x_iによって説明する回帰モデル$y_i = \alpha + \beta x_i + \varepsilon_i$を考え、

$$x_i = \begin{cases} 1 & \text{中卒} \\ 2 & \text{高卒} \\ 3 & \text{大卒} \end{cases}$$

のように数値化したとしよう。そうすると、

中卒　　$\mathrm{E}(y_i \mid x_i = 1) = \alpha + \beta$
高卒　　$\mathrm{E}(y_i \mid x_i = 2) = \alpha + 2\beta$
大卒　　$\mathrm{E}(y_i \mid x_i = 3) = \alpha + 3\beta$

となって、大卒の限界賃金が中卒のそれの３倍であることを暗黙に仮定していることになってしまう。

このようなモデル化（パラメータの立て方）は果たして妥当であろうか。

ダミー変数

このような時に、よく用いられるのがダミー変数である。ダミー変数は、観測された固体がある属性を持っているときに１、そうでないときに０という値をもつ変数のことである。ある意味で、属性を示すフラグ（旗）変数である。

先ほどの例では、

$$d_{1i} = \begin{cases} 1 & \text{高卒} \\ 0 & \text{それ以外} \end{cases} \qquad d_{2i} = \begin{cases} 1 & \text{大卒} \\ 0 & \text{それ以外} \end{cases}$$

という２つのダミー変数を導入することによって、

$$y_i = \alpha + \beta_1 d_{1i} + \beta_2 d_{2i} + \varepsilon_i$$

というモデルを考える。そうすると、

中卒　$\mathrm{E}(y_i \mid d_{1i}=0, d_{2i}=0)=\alpha$

高卒　$\mathrm{E}(y_i \mid d_{1i}=1, d_{2i}=0)=\alpha+\beta_1$

大卒　$\mathrm{E}(y_i \mid d_{1i}=0, d_{2i}=1)=\alpha+\beta_2$

とパラメータ化されていることになる。このとき、β_1は中卒を基準とした高卒の賃金プレミアムに相当するし、β_2 は中卒を基準とした大卒の賃金プレミアムに相当する。

だだ、ダミーの立て方には注意が必要である。定数項のある回帰式の場合、ダミーは（属性数－１）個までしか入れることが出来ない。今の学歴の例では、説明変数行列 $\mathbf{X}$ は

$$\mathbf{X}=\begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{pmatrix}$$

となっているので、もしも中卒ダミーd_{0i}を入れてしまうと、

$$\tilde{\mathbf{d}}_0+\tilde{\mathbf{d}}_1+\tilde{\mathbf{d}}_2=\tilde{\mathbf{1}}$$

となって、$\mathbf{X}$ を構成する列ベクトルが一次独立ではなくなってしまう。（従って $\mathbf{X}'\mathbf{X}$ の逆行列は存在しない）。

応用例

①二群の平均の差の検定

$$d_i=\begin{cases}1 & \text{Group A}\\ 0 & \text{Group B}\end{cases}$$

というダミー変数を用いると、二群の分散が等しい時の平均の差の検定が、回帰モデル $y_i=\alpha+\beta\, d_i+\varepsilon_i$　で行なえる。

A群　$\mathrm{E}(y_i \mid d_i=1)=\alpha+\beta=\mu_A$

B群　$\mathrm{E}(y_i \mid d_i=0)=\alpha=\mu_B$

であるから、

$$\begin{cases}H_0:\beta=0\\ H_1:\beta\neq 0\end{cases}\Leftrightarrow\begin{cases}H_0:\mu_A=\mu_B\\ H_1:\mu_A\neq\mu_B\end{cases}$$

となって、パラメータβの検定問題に帰着できる。

②季節性の除去

季節ダミーを導入することによって、簡便に季節性を除去することが出来る。但し、この場合も、定数項のある回帰モデルでは（季節数－１）しかダミー変数は入れられないので、注意。

ダミー変数の高度な使い方

いままで、平均水準の違いをダミー変数を用いてモデル化する方法について見てきた。次に、反応（回帰係数）の異なるモデルをダミー変数を用いて表現することを考えよう。

バブル期とデフレ期で限界消費性向の異なるケインズ型消費関数を考えよう。

バブル期　　$C_i = \alpha + \beta_b Y_i + \varepsilon_i$

デフレ期　　$C_i = \alpha + \beta_d Y_i + \varepsilon_i$

これをバブルダミー、

$$d_i = \begin{cases} 1 & \text{bubble} \\ 0 & \text{deflation} \end{cases}$$

を用いて、

$$C_i = \alpha + \beta_0 (Y_i \times d_i) + \beta_1 Y_i + \varepsilon_i$$

とモデル化しよう。このとき、

bubble　　$\mathrm{E}(y_i \mid d_i = 1) = \alpha + (\beta_0 + \beta_1) Y_i$

deflation　　$\mathrm{E}(y_i \mid d_i = 0) = \alpha + \beta_1 Y_i$

となるので、

$$\begin{cases} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{cases} \Leftrightarrow \begin{cases} H_0 : \beta_b = \beta_d \\ H_1 : \beta_b \neq \beta_d \end{cases}$$

となって、パラメータ β_0 の検定問題に帰着できる。

構造変化の検定（Chow Test）

今までの例では、構造の違いは１つのパラメータだけであった。では、２つ以上のパラメータが異なるときは、どのように検定すれば良いのだろうか。→Ｆ検定。

例として、資産効果を考慮した消費関数、

$$C_i = \alpha + \beta_1 Y_i + \beta_2 W_i + \varepsilon_i \qquad \cdots \quad ①$$

を考えることにしよう。ここでの仮説は、

仮説：　バブル期とデフレ期では消費性向は異なる。

である。そこで、先ほどのバブルダミーを導入して制約のない回帰モデルを推定する。

$$C_i = \alpha + \beta_1 Y_i + \delta_1 (Y_i \times d_i) + \beta_2 W_i + \delta_2 (W_i \times d_i) + \varepsilon_i \qquad \cdots \quad ②$$

すると仮説は、

$$\begin{cases} H_0 : \delta_1 = \delta_2 = 0 \\ H_1 : \text{ otherwise} \end{cases}$$

のように書きかえることが出来る。これに基づく F 検定の具体的な手順は、

(a) 帰無仮説 H_0 すなわち①式を OLS 推定し、その残差２乗和（RRSS と呼ぶ）を求める。

(b) 次に対立仮説 H_1 すなわち②式を OLS 推定し、その残差２乗和（URSS と呼ぶ）を求める。

(c) 統計量

$$f = \frac{(RRSS - URSS)/2}{URSS/(n-5)} \sim F(2, n-5)$$

を求めると、帰無仮説の下で f は分子自由度２、分母自由度ｎ－５のＦ分布に従うので、Ｆ分布表を用いて検定する。

となる。

§6.2. 説明変数が質的データである回帰モデル

9 月 10 日授業で説明したので省略。教科書 10 章（pp.195―201）とほぼ同内容。

なお、教科書 p.200 にある(10-3-10)式が証明なしに提示されているが、この導出は、Taylor 展開とロジスティック関数の微分が必要になり、やや高度なので、この授業では説明しないことにする。