非市場評価研究の実践

補章　回帰分析の基礎

立命館大学経済学部

寺　脇　　拓

1. 最小二乗法

1.1 回帰分析

- **回帰分析**(regression analysis)
  - ある変量(y)が他の変量群($x_1, x_2, \cdots, x_n$)によってどのように説明されるかを統計学的に分析する手法の総称。
- **線形回帰モデル**(linear regression model)
  - 回帰分析において最も基本的なモデル。
  - **単回帰モデル**：ある変量(y)を一つの変量(x)に回帰させる。

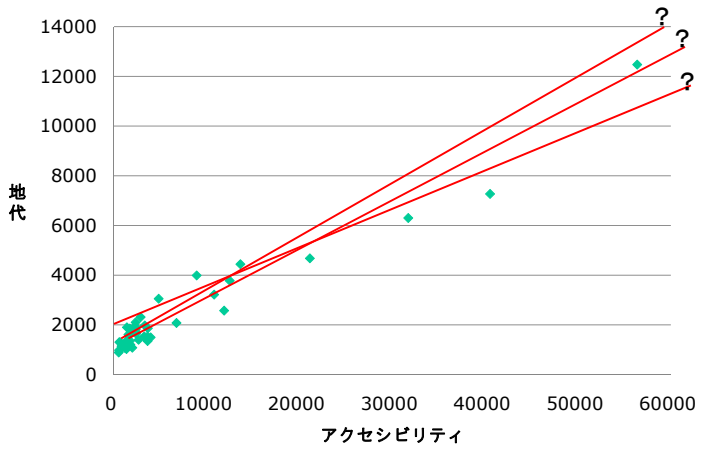
y：被説明変数(従属変数)　β：係数パラメータ

$$y = \alpha + \beta x + \epsilon$$

ε：誤差項(残差)　α：定数項　x：説明変数(独立変数)

  - **重回帰モデル**：ある変量(y)を複数の変量群で構成されるベクトル($x_1, x_2, \cdots, x_n$)に回帰させる。

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n + \epsilon := \boldsymbol{\beta}' \boldsymbol{x} + \epsilon$$



図A.1　直線の当てはめ

1.2 最小二乗推定

- アクセシビリティx (交通条件の良さ)と地代y (地価を年当たりの価値に直したもの)について、三つ地点のデータ(x_1, y_1)、(x_2, y_2)、(x_3, y_3)が得られている。
- アクセシビリティと地代の間にある真の関係が次の線形式で表わされるとする。

$$y = \alpha + \beta x \quad \cdots(1)$$

- 通常(x_1, y_1)、(x_2, y_2)、(x_3, y_3)は(1)式を満たさないが、そのずれをϵ で表して(1)式の右辺に加えてやれば、(x_1, y_1, ϵ_1)、(x_2, y_2, ϵ_2)、(x_3, y_3, ϵ_3)は次式を満たす。

$$y = \alpha + \beta x + \epsilon \quad \cdots(2)$$

- ここで、(2)式に(x_1, y_1, ϵ_1)、(x_2, y_2, ϵ_2)、(x_3, y_3, ϵ_3)を代入したものをすべて「$\epsilon =$ 」の形に直す。

$$\epsilon_1 = y_1 - \alpha - \beta x_1$$
$$\epsilon_2 = y_2 - \alpha - \beta x_2$$
$$\epsilon_3 = y_3 - \alpha - \beta x_3$$

- それぞれ両辺を二乗して合計する。
  - これを**残差平方和** (sum of squared errors : SSE)という。

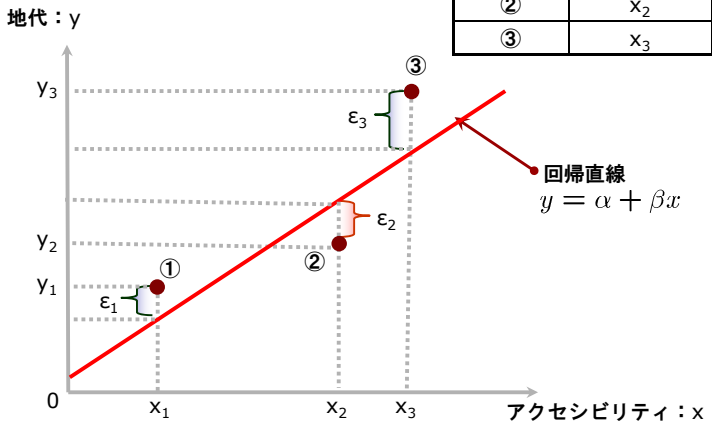
$$SSE = \sum_{i=1}^{3} \epsilon_i^2 = \sum_{i=1}^{3} (y_i - \alpha - \beta x_i)^2$$

- 残差平方和を最小にするようなαとβを計算する。
  - SSEはαとβの二次関数であるので、これをαとβそれぞれで偏微分し、イコール0とした連立方程式を解くことによって、それらの推定値が得られる。

$$\frac{\partial SSE}{\partial \alpha} = 0, \quad \frac{\partial SSE}{\partial \beta} = 0 \quad \cdots(3)$$

- この方法を**最小二乗法**(ordinary least squares : OLS)という。

ID	アクセシビリティ	地代
①	x_1	y_1
②	x_2	y_2
③	x_3	y_3



図A.2　回帰直線と誤差

1.3 最小二乗推定量

- (3)式で表される連立方程式を解くと次式が得られる。

$$\hat{\beta} = \frac{\sum_{i=1}^{3}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{3}(x_i - \bar{x})^2}, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

  - ここでのαとβは、真のパラメータではなく、推定される値を意味するため「^(ハット)」の記号が付けられている。
  - 変数の上についている「−(バー)」はその変数の平均を意味する。
  - 観測値数がnの場合は「3」のところが「n」に置き換えられる。
  - 二つ目の式からわかるように、xとyの平均は必ず推定された回帰直線を通る。
- これらを**最小二乗推定量** (OLS estimator)という。
  - **推定量**(estimator)は「推定の仕方」を意味し、**推定値**(estimate)はその推定の仕方に従って計算された値を意味する。

2. 分散分析と決定係数

2.1 分散分析

- 回帰による個人iのyの**予測値**($\hat{y}_i$)は次式で表される。

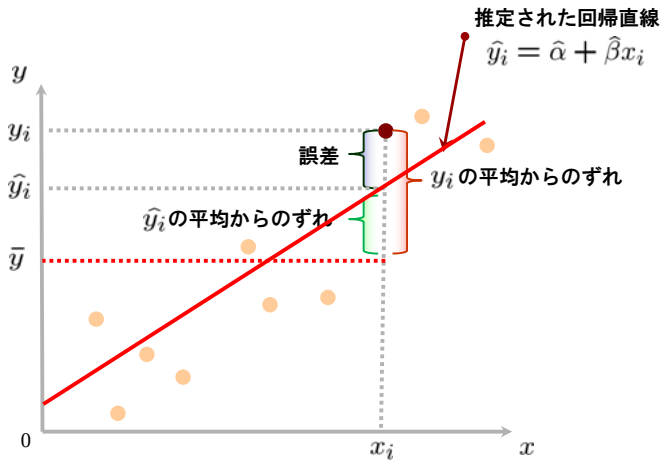
$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

- 実測値yの**全変動**(total sum of squares: SST)は、**回帰による変動**(予測値の変動 regression sum of squares: SSR)と**誤差による変動**(residual sum of squares: SSE)に分けられる。
  - **変動**(variation)とは、ある変数の観測値とその平均までの距離の二乗和で表わされる測度を意味する。

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2,\ SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2,\ SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SST = SSR + SSE$$

- この分解により全変動のうちどれくらいが回帰によって説明できるかを分析することを**分散分析**(analysis of variance: ANOVA)という。



図A.3　予測値と誤差

2.2 決定係数

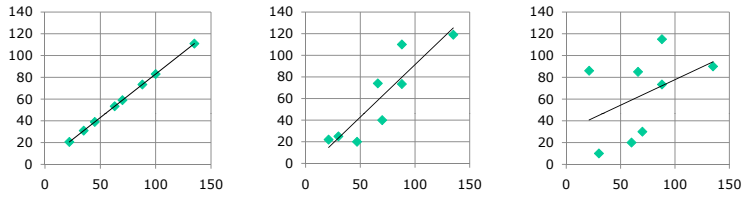
- 回帰による変動(SSR)が全変動(SST)に占める割合を**決定係数**といい、**R^2**で表わされる。

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- 観測値の回帰直線への当てはまりがよいほど決定係数は1に近づく。
- ただし重回帰モデルにおいて、説明変数の数を増やすほど決定係数は高くなるため、一般には残差の自由度を考慮した**自由度調整(修正)済み決定係数**(adjusted R^2)が用いられる。

$$\bar{R}^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

  - k：説明変数の数。



図A.4　決定係数と直線の適合度

3. 係数に関する仮説検定

3.1 B君は立命生？

- A君は、最近知り合ったB君がどうも立命生ではないように思えて仕方ない。けれども面と向かってそれを聞くのも気が引ける。そこでA君は、とりあえず「B君が立命生である」と仮定して、普通に接してみることにした。そしてそのときにもしA君が立命生としてありえない行動をとったとしたら、「B君は立命生ではない」と判断しようと考えた。小雨が降るある日、A君はコラーニングでB君と出会う。

A 「おう、B。これからリンクで一緒に飯食わん？」

B 「ええよ。俺も今しか時間ないわ。」

A 「よっしゃ、ほなら･･･あっ、しもた。教室に傘忘れてきたわ。悪いけど先リンク行ってて。」

B 「えっ、どこ？」

A 「リ・ン・ク。ほな後で。」

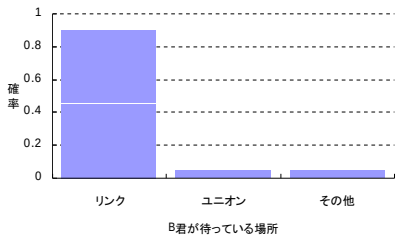
B 「えっ、ちょっ･･･。」

- このとき、A君は本当に傘を教室に忘れてきていたのであって、決してB君を試そうとしたのではない。しかし教室に向かう途中でA君は気づいた。

A 「(これはいい機会かもしれん･･･。)」

- 傘を回収し、A君はリンクに向かった。彼は考えていた。

A 「(あれだけなんども「リンク」言うたんやから、もしBが立命生なら、かなりの確率･･･そうやな90%ぐらいの確率でリンクで待っているはずや。ちゃんと聞き取れてへんかったとしても、これから飯食おういうてんねんから、最悪「ユニオン」にはいるやろ。この確率は5%ぐらいかな。それ以外の場所にいるのが残りの5%やな。Cキューブは休みやし、この状況で、リンク、ユニオン以外の場所にいくなんてことは、立命生やったらありえへんわ。もしそんなありえへんようなことが起こったとしたら、そのときは、Bは立命生やないちゅうことになる･･･)」



- リンクにたどり着いたA君。しかし、B君は見当たらない。一抹の不安を覚えながら、A君はユニオンに向かった。雨が冷たく感じられる。

B 「おう、A！」

A 「あっ･･･B！」

B 「ごめん、リンク言うてたんやったっけ。なんかいっぱいやったからこっちにきてん。」

A 「ああ･･･そっか。」

B 「席とってあんで。こっちや。」

A 「(ああ、よかった･･･。)」

- A君は素直にほっとした。しかし同時に、なにか釈然としない感覚がこみ上げてくる。

A 「(自分の予測では、Bがリンクで待ってへん確率はたった10%や。これはありえへん結果とちゃうやろか･･･。もし「ありえへん」の意味を「10%」にまで引き上げてたら、「Bは立命生やない」ちゅう結論になってたんや･･･。)」

A 「ま、えっか。」

B 「どないしてん急に。」

A 「どない頑張ったって真実はわからへんねんから、いまある情報だけで判断せなあかんちゅうこっちゃ。」

B 「はあ？自分、頭大丈夫か？」

- B君のつっこみにA君は笑顔で返した。いつしか雨はあがっていた。

3.2 仮説検定の考え方

- **統計的仮説検定**(statistical hypothesis testing)
  - 仮説が正しい場合には、ある値が実現するのはめったに起きないという事実を利用して、その事実が起きたことを理由にその仮説を否定し、対立する仮説を採択する論証法。
- **帰無仮説**(null hypothesis)
  - 検定される仮説(B君は立命生である)。
- **対立仮説**(alternative hypothesis)
  - 帰無仮説が正しくないと判断される場合に採択される仮説(B君は立命生でない)。

> 帰無仮説が正しいとしたときに、ある確率的に変動する量(**検定統計量** test statistic)がどのような分布をもつかを考え、観測値から計算されるその値がどれくらいの確率で起こりえるものなのかをみることで、その帰無仮説の妥当性を見極める。

3.3 仮説検定の一般的な手続き

- 帰無仮説と対立仮説を立てる。
- 帰無仮説が正しいときの検定統計量の分布を導く。
- 「めったに起こらない」という言葉が表す小さな確率(**有意水準** significance level)をきめる。
- この確率に基づいて、帰無仮説を選ぶか、対立仮説を選ぶかを判断する検定統計量の基準値(**臨界値** critical value)を導く。
  - 臨界値を基準にして対立仮説を支持する領域を**棄却域**(critical region)という。
- 観測値を使って、検定統計量の実現値を計算する。
- この値が棄却域に含まれるかどうかをみて、棄却域に含まれれば帰無仮説を**棄却**(reject)し、対立仮説を**採択**(accept)する。

3.4 係数の有意性検定

- (2)式の回帰モデルにおいて ϵ が正規分布に従うと仮定すると、次の t は自由度 $n-2$ (n は観測値数)のt分布に従うことが知られている。

$$t = \frac{\hat{\beta} - \beta}{SE(\hat{\beta})}$$

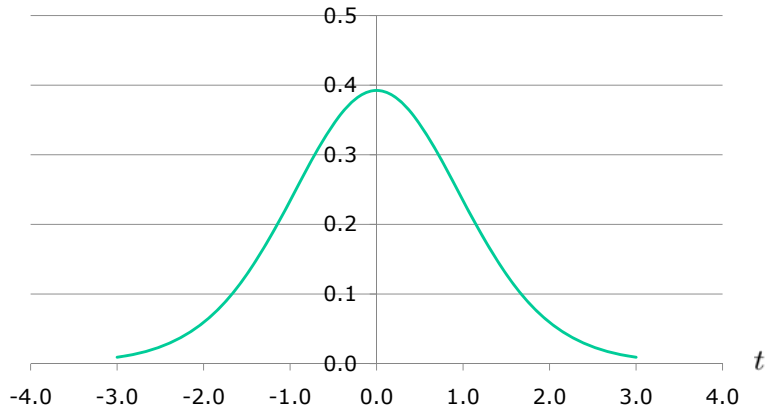
  - $SE(\hat{\beta})$ は係数推定値の標準誤差(標準偏差の推定値)を表す。
- ここで帰無仮説を $\beta = 0$ 、対立仮説を $\beta \neq 0$ とする。
- 帰無仮説が正しいとき、次の t は自由度 $n-2$ のt分布に従うため、これを検定統計量として用いる。

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})} \quad \cdots(4)$$

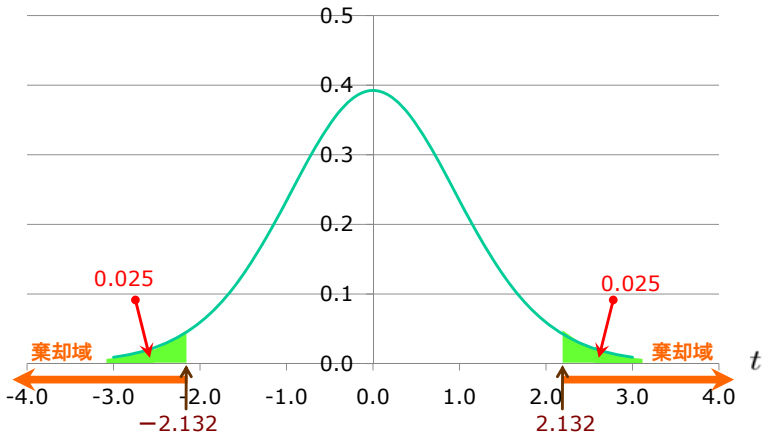
- これを**t値**(t-value)という。

■例

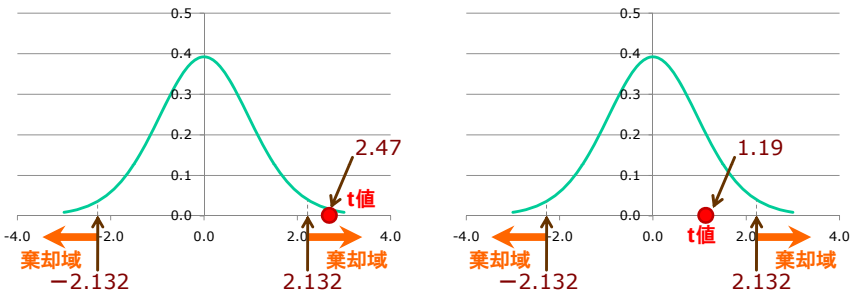
- 観測値数が17のとき(4)式の t は自由度15のt分布に従う。
- 有意水準を5%に設定すると、棄却域(対立仮説を支持する領域)は、 $t < -2.131$、$t > 2.131$ となる。
- 実際のt値を計算し(コンピュータが計算する)、これが棄却域に含まれれば、帰無仮説 $\beta = 0$ を棄却し、対立仮説 $\beta \neq 0$ を採択する。
- このとき、係数パラメータは5%水準で有意である(あるいはゼロと有意差がある)という。
  - 説明変数 x は被説明変数 y に有意に影響を与えるといえる。



図A.5　自由度15（観測値数が17の場合）のt分布



図A.6　有意水準を5%とした場合の棄却域



a. $\beta = 0$が棄却される場合　　b. $\beta = 0$が棄却できない場合

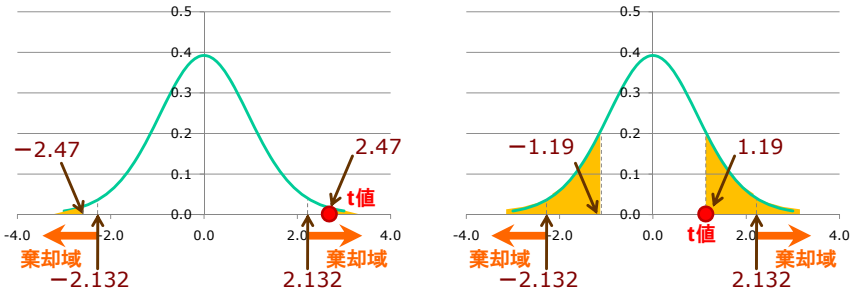
図A.7　t値の位置と検定結果

3.5 p値による検定結果の見極め

- 得られたt値がt_0で表されるとき、次の確率はp値(p-value)と呼ばれる。

$$p := P(t < -t_0, t > t_0)$$

- t値が有意水準αの棄却域に入るとき、p値はαよりも小さくなり、入らないときには、p値はαよりも大きくなる。
- p値を示せば、それだけである特定の有意水準で$\beta = 0$が棄却できるかどうかがわかる。



a. $\beta = 0$が棄却される場合　　b. $\beta = 0$が棄却できない場合

図A.8　t値とp値