重回帰分析

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$

4. その他の問題

Ch.6 重回帰分析：その他の問題

1. データ単位について
2. 関数形
3. 決定係数と説明変数の選択
4. 予測と残差分析*

6.1 データ単位について

- 変数の再定義
  - 変数の単位変更による影響
    - 係数、標準誤差
      - x_j をc倍すると、その係数は$1/c$に
      - y をc倍すると、**すべての** OLS係数はc倍に
  - 影響を受けないもの
    - 有意性、解釈
      - いかなる変数の単位変更でも、t 値や F 値に影響なし
    - 対数形
      - 対数関数は、切片以外の係数への影響なし

→ table 6.1

TABLE 6.1

Effects of Data Scaling

Dependent Variable	(1) *bwght*	(2) *bwghtlbs* = *bwght*/16	(3) *bwght*
Independent Variables			
cigs	−.4634 (.0916)	−.0289 (.0057)	——
packs	——	——	−9.268 (1.832)
faminc	.0927 (.0292)	.0058 (.0018)	.0927 (.0292)
intercept	116.974 (1.049)	7.3109 (.0656)	116.974 (1.049)
Observations	1,388	1,388	1,388
R-Squared	.0298	.0298	.0298
SSR	557,485.51	2,177.6778	557,485.51
SER	20.063	1.2539	20.063

1/16

TABLE 6.1

Effects of Data Scaling

Dependent Variable	(1) *bwght*	(2) *bwghtlbs* = *bwght*/16	(3) *bwght*
Independent Variables			
cigs	−.4634 (.0916)	−.0289 (.0057)	——
packs = *cigs*/20	——	——	−9.268 (1.832)
faminc	.0927 (.0292)	.0058 (.0018)	.0927 (.0292)
intercept	116.974 (1.049)	7.3109 (.0656)	116.974 (1.049)
Observations	1,388	1,388	1,388
R-Squared	.0298	.0298	.0298
SSR	557,485.51	2,177.6778	557,485.51
SER	20.063	1.2539	20.063

×20

6.1 データ単位について

Dependent / Independent	y	cy	y
x_1	β_1 (se_1)	$c\beta_1$ (c^*se_1)	——
dx_1	——	——	β_1/d (se_1/d)
x_2	β_2 (se_2)	$c\beta_2$ (c^*se_2)	β_2 (se_2)
Intercept	β_0 (se_0)	$c\beta_0$ (c^*se_0)	β_0 (se_0)
R-squared	R^2	R^2	R^2
SSR	SSR	c^2*SSR	SSR

Standard errors in parentheses

6.1 データ単位について

- 6-1a 標準化係数Beta Coefficients
  - 全変数の標準化
    - 各変数をその平均で引き、標準偏差で割る
  - 標準化係数の意味
    - xの1標準偏差変化に対するyの標準偏差変化を反映
    - 係数推定値を比較し、「どの変数が最も重要か」等の考察が可能
    - 標準化変数と非標準化変数のいずれの回帰結果も、統計的有意性は同じ

6-2 関数形について

- 6-2a 対数形
  - メリット
    - 変化率や弾力性の解釈に便利
      - 対数変数の係数は単位変更に不変
    - 対数変換は、外れ値の問題を排除・緩和
      - 正規性と均一分散の仮定に貢献
  - 注意点
    - 変数に非正値を含む場合、対数変換できない
      - 年数や%の単位の変数は、通常、対数変換しない
      - 予測の際に対数変換を元に戻すことは難しい

6-2 関数形について

- 対数形の解釈
  - $\ln(y) = \beta_0 + \beta_1 \ln(x) + u$
    - β_1はxに対するyの弾力性
    - β_1 is the elasticity of y with respect to x.
  - $\ln(y) = \beta_0 + \beta_1 x + u$
    - β_1はxが1単位変化したときのyの（近似）変化率
    - β_1 is approximately the percentage change in y given a 1 unit change in x.
  - $y = \beta_0 + \beta_1 \ln(x) + u$
    - β_1はxの変化率が変化したときのyの（近似）単位変化
    - β_1 is approximately the change in y for a 100 percent change in x.

6-2 関数形について

Model	Dependent Variable	Independent Variable	Interpretation of β_1
level-level	y	x	$\partial y/\partial x$
level-log	y	$\log(x)$	$\partial y/(\partial x/x)$
log-level	$\log(y)$	x	$(\partial y/y)/\partial x$
log-log	$\log(y)$	$\log(x)$	$(\partial y/y)/(\partial x/x)$

6-2 関数形について

- 6-2b 2次関数形
  - 例：賃金方程式

$$\widehat{wage} = \underset{(.35)}{3.73} + \underset{(.041)}{.298}\, exper - \underset{(.0009)}{.0061}\, exper^2$$

$n = 526, R^2 = .093$

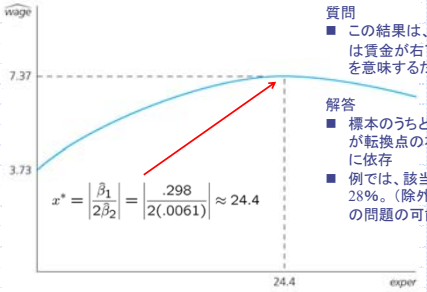
    - 職業経験の凹関数（下に凹）
  - 職業経験の限界効果

$$\frac{\Delta wage}{\Delta exper} = .298 - 2(.0061)exper$$

    - 最初の1年間の経験後、2年目の賃金が約$0.30上昇
      ← 0.298 – 2(0.0061)(1) = $0.29

6-2 関数形について

- 最高賃金時の職歴



質問
- この結果は、職歴24.4年以後は賃金が右下がりになることを意味するだろうか？

解答
- 標本のうちどれほどの観測数が転換点の右側に位置するかに依存
- 例では、該当は観測数の約28%。（除外変数等）特定化の問題の可能性

6-2 関数形について

◆例6.2：住宅価格への大気汚染の影響

$$\widehat{\log(price)} = \underset{(.57)}{13.39} - \underset{(.115)}{.902}\log(nox) - \underset{(.043)}{.087}\log(dist) - \underset{(.165)}{.545}\,rooms + \underset{(.013)}{.062}\,rooms^2 - \underset{(.006)}{.048}\,stratio$$

$n = 506, R^2 = .603$

- 制御変数：空気中の窒素酸化物、雇用センターからの距離、平均的な学生/教師の比率

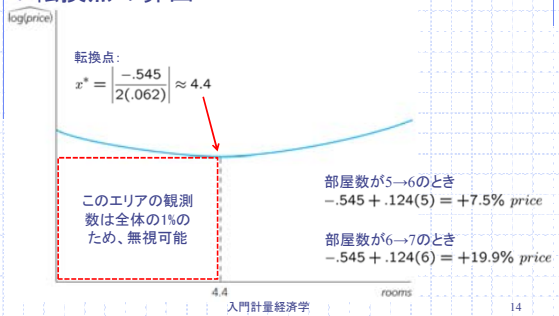
■ 部屋数の影響

$$\Rightarrow \frac{\Delta \log(price)}{\Delta rooms} = \frac{\%\Delta price}{\Delta rooms} = -.545 + .124 rooms$$

- この結果は、「部屋数の少ない家では、部屋数増加がより低い住宅価格と関連している」ことを意味するのか？

6-2 関数形について

◆転換点の算出



6-2 関数形について

◆他の可能性

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 [\log(nox)]^2 + \beta_3 crime + \beta_4 rooms + \beta_5 rooms^2 + \beta_6 stratio + u$$

$$\Rightarrow \frac{\Delta \log(price)}{\Delta \log(nox)} = \frac{\%\Delta price}{\%\Delta nox} = \beta_1 + 2\beta_2 [\log(nox)]$$

◆多項式

$$cost = \beta_0 + \beta_1 quantity + \beta_2 quantity^2 + \beta_3 quantity^3 + u$$

6-2 関数形について

◆6-2c 交差項モデル

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + \underbrace{\beta_3 sqrft \cdot bdrms}_{\text{交差項}} + \beta_4 bthrms + u$$

$$\Rightarrow \frac{\Delta price}{\Delta bdrms} = \beta_2 + \beta_3 sqrft$$

- 部屋数の効果は面積の水準による

■ 交差項効果による係数解釈に要注意

- β_2 = 家の面積がゼロ(!)のときの部屋数効果
  →再係数化

6-2 関数形について

◆交差的効果の再係数化Reparametrization

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u$$

$$y = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \beta_3 (x_1 - \mu_1)(x_2 - \mu_2) + u$$

- μ_1, μ_2：母集団でのx_1, x_2の平均←標本平均で代替
- δ_2：全変数が平均値であるときのx_2の効果

■ 再係数化のメリット

- 全係数の解釈が容易
- 平均値での部分効果の標準誤差算出が容易
- 必要に応じて、交差項は（平均値の代わりに）関心ある他の数値でも計算可能

6-2 関数形について

◆6-2d 平均部分効果の計算

- 2次関数、交差項、非線形のモデルでは、部分効果は説明変数の値に依存
- 平均部分効果APEは、被説明変数と各説明変数との間の関係を記述するための要約的指標
- 部分効果を計算し、推定係数を代入後、標本全体の部分効果を平均化

6-3 決定係数と説明変数選択

◆決定係数R^2について
- 高いR^2は必ずしも因果的解釈を意味しない
- 低いR^2は部分効果の正確な推定を排除するものではない

◆母集団R^2と標本R^2
- 母集団R^2 $1-\frac{\sigma_u^2}{\sigma_y^2}$ ← $\frac{u\text{の分散}}{y\text{の分散}}$ 本来求めるべきR^2
- 標本R^2 $R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{(SSR/n)}{(SST/n)}$

6-3 決定係数と説明変数選択

◆6-3a 自由度修正済み決定係数Adjusted R^2
- 自由度を考慮した推定量

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

  - 分子と分母の自由度を修正
  - 説明変数の追加ごとにペナルティを課す
  - 追加変数のt値が$|t|>1$の場合のみ増加
- 通常R^2と修正R^2の関係

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

  - 修正R^2はマイナスの値になることも

6-3 決定係数と説明変数選択

◆6-3b 修正R^2によるNonnestedモデルの選択
- 非入れ子型Nonnestedモデル
  - いずれのモデルも他方の特殊ケースでないモデル
  1. $rdintens = \beta_0 + \beta_1 \log(sales) + u$ $R^2 = .061, \bar{R}^2 = .030$
  2. $rdintens = \beta_0 + \beta_1 sales + \beta_2 sales^2 + u$ $R^2 = .148, \bar{R}^2 = .090$
  - 両モデルのR^2の比較は、モデル1にとって係数の少なさのために不公平
  - 例では、自由度の差を調整した後でさえモデル2の方が良好なフィットを明示

6-3 決定係数と説明変数選択

◆被説明変数が異なるモデルの比較
- 被説明変数が異なる場合、通常R^2と修正R^2を使用してモデルを比較できない
  - 例6.4：CEO報酬と企業業績

$$\widehat{salary} = 830.63 + .0163\ sales + 19.03\ roe$$
$$(223.90)\ \ (.0089)\ \ (11.08)$$
$$n = 209, R^2 = .029, \bar{R}^2 = .020, SST = 391,732,982$$

$$\widehat{lsalary} = 4.36 + .275\ lsales + .0179\ roe$$
$$(0.29)\ \ (.033)\ \ (.0040)$$
$$n = 209, R^2 = .282, \bar{R}^2 = .275, SST = 66.72$$

  - モデルで説明される対数値の変動はレベルよりも少ない

6-3 決定係数と説明変数選択

◆6-3c 多すぎる要因の制御
- 場合によっては特定の変数を固定すべきでない
  - 交通事故の各州ビール税（および他の要因）への回帰で、ビール消費を直接的に制御すべきではない
  - 各家庭医療費の農家間での農薬使用への回帰では、医師の診察を制御すべきではない
- 異なる回帰式は異なる目的を提供
  - 住宅価格の家の特性への回帰では、回帰の目的が査定の有効性調査であるなら価格評価を含めるが、そうでなければ含むべきではない

6-3 決定係数と説明変数選択

◆6-3d 誤差分散を減らす説明変数の追加
- 説明変数の追加は、誤差分散を減少
  - 一方、多重共線性の問題を悪化させることがある
  - ただし他の説明変数と無相関ならば、多重共線性を増やさずに誤差分散を減らすので追加すべき
  - しかし、このような無相関変数は見つけにくい
- 例：個人ビール消費とビール価格
  - ビール消費のその価格への回帰式において、各個人特性を説明変数として含むことは、価格弾力性をより正確に推定

6-4 予測と残差分析

◆6-4c log(y)が被説明変数のときのyの予測

- 指数化（逆対数化）

$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$

$\Rightarrow \quad y = \exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)\exp(u)$

- 「誤差項uは説明変数xと独立」の仮定

$\Rightarrow \quad E(y|\mathbf{x}) = \exp(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)E(\exp(u))$

$\Rightarrow \quad \hat{y} = \exp(\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2+\cdots+\hat{\beta}_k x_k)\left(\frac{1}{n}\sum_{i=1}^{n}\exp(\hat{u}_i)\right)$

→yの予測値の算出

6-4 予測と残差分析

◆対数モデルと非対数モデルのR^2

$\widehat{salary} = 613.43 + .0190\ sales + .0234\ mktval + 12.70\ ceoten$
$(65.23)\ (.0100)\ (.0095)\ (5.61)$

$n = 177, R^2 = .201$

$\widehat{lsalary} = 4.504 + .163\ lsales + .109\ mktval + .0117\ ceoten$
$(.257)\ (.039)\ (.050)\ (.0053)$

$n = 177, R^2 = .318 \quad \bar{R}^2 = .243$

- 非対数給与変数の予測のためのR^2
  - モデル2の回帰式では本来の対数給与のR^2を修正
  - これにより両モデルのR^2を直接比較可能に