回帰分析I

3. データセットの種類、よく使う数学、係数の読み方

データセットの種類

- イントロダクションで見たように、実証分析はデータを使います。
- ここではデータの種類について説明します。
- 大きく分けると3種類。
  - 横断面（cross section）データ
  - 時系列（time series）データ
  - パネル（panel）データ
- それぞれ見ていきましょう。

横断面データ

- 横断面データは、ある一時期における<u>異なる主体</u>に関するデータ。
- 「主体」は、個人、企業、県、国など、分析の対象によって異なります。
- 前回扱ったカリフォルニア州の学区ごとのテスト成績データは、クロスセクションデータの一例。
  - ある一つの時期（1998年）における420個の主体（学区）
- 一般的に、観察される主体の数は N で表されます。
  - カリフォルニア州のデータでは $N = 420$

横断面データ

ID	dist_cod	county	district	testscr	str	avginc
1	75119	Alameda	Sunol Glen Unified	690.80	17.89	22.69
2	61499	Butte	Manzanita Elementary	661.20	21.52	9.82
3	61549	Butte	Thermalito Union Elementary	643.60	18.70	8.98
4	61457	Butte	Golden Feather Union Elementary	647.70	17.36	8.98
5	61523	Butte	Palermo Union Elementary	640.85	18.67	9.08
6	62042	Fresno	Burrel Union Elementary	605.55	21.41	10.41
7	68536	San Joaquin	Holt Union Elementary	606.75	19.50	6.58
8	63834	Kern	Vineland Elementary	609.00	20.89	8.17
9	62331	Fresno	Orange Center Elementary	612.50	19.95	7.39
10	67306	Sacramento	Del Paso Heights Elementary	612.65	20.81	11.61

ソース：Stock and Watson (2007) *Introduction to Econometrics*, Pearson Education, Inc

- 第一行目は変数名。
- 行の順番は任意で、ID（一列目）はデータ整備のために任意に割り当てられた番号にすぎない。
- 横断面データの分析は、ある一時期における主体の変数の大きさの違いを使って、変数間の関係を明らかにしようというもの。

時系列データ

- 時系列データは<u>一つの主体</u>に関するデータが、多くの期間に渡って集められたもの。
  - 集められる間隔は、データによる。
  - 例えば、一日おき（日次データ）、一か月おき（月次データ）、四半期おき（四半期データ）、一年おき（年次データ）など。
  - マクロ・データ（GDP、マネーサプライ、物価など）であることが多い。
- 一般的に、観察される時点の数は T で表される。
  - 例えば、年次データで10年あれば、$T = 10$
  - 例えば、四半期データで10四半期あれば、$T = 10$
  - 例えば、月次データで10か月あれば、 $T = 10$
  - 例えば、日次データで10日あれば、$T = 10$
- 行の順番（並び方）に意味あり。
  - 前の行は後の行よりも時間的に先のもの。

5

時系列データ

東証一部　Klab（証券コード：3656）の株価・出来高の時系列データ

日付	始値	高値	安値	終値	出来高	調整後終値
2018年3月29日	1,700	1,732	1,679	1,708	1,482,200	1,708
2018年3月28日	1,613	1,675	1,613	1,673	746,200	1,673
2018年3月27日	1,642	1,673	1,630	1,638	820,300	1,638
2018年3月26日	1,600	1,612	1,530	1,606	1,292,000	1,606
2018年3月23日	1,600	1,637	1,594	1,618	1,530,700	1,618
2018年3月22日	1,683	1,710	1,678	1,682	1,047,600	1,682
2018年3月20日	1,780	1,815	1,673	1,681	3,518,800	1,681

（ソース：https://stocks.finance.yahoo.co.jp/stocks/history/?code=3656.T）

- 行の順番（並び方）に意味あり。
  - 前の行は後の行よりも時間的に先のもの。
- 時系列データを使った分析は、時間を通じた変数の動きから変数間の関係を明らかにしようとするもの。

6

パネルデータ

- パネルデータは、複数の主体に関するデータで、各主体について2つ以上の時期において観察されたデータ。
- 一般的に、観察される主体の数はNで、観察される時点の数はTで表されます。
- 例：都道府県それぞれについて、1980年、1981年、1982年のデータがある → $N = 47$, $T = 3$

7

id	t	wage	wks	fem	union	ed	exp	ms
1	1	260	32	0	0	[illegible]	[illegible]	[illegible]
1	2	305	43	0	0	[illegible]	[illegible]	[illegible]
1	3	402	40	0	0	[illegible]	[illegible]	[illegible]
1	4	402	39	0	0	[illegible]	[illegible]	[illegible]
1	5	429	42	0	0	9	7	1
1	6	480	35	0	0	9	8	1
1	7	[illegible]	[illegible]	[illegible]	0	9	9	1
2	1	[illegible]	[illegible]	[illegible]	0	11	30	1
2	2	[illegible]	[illegible]	[illegible]	0	11	31	1
2	3	[illegible]	[illegible]	[illegible]	1	11	32	1
2	4	695	30	0	0	11	33	1
2	5	810	30	0	0	11	34	1
2	6	890	37	0	0	11	35	1
2	7	912	30	0	0	11	36	1
3	1	285	50	0	1	12	6	1
3	2	624	51	0	1	12	7	1
3	3	698	50	0	1	12	8	1
3	4	737	52	[illegible]	[illegible]	[illegible]	[illegible]	1
3	5	809	52	[illegible]	[illegible]	[illegible]	[illegible]	1
3	6	879	52	[illegible]	[illegible]	[illegible]	[illegible]	0
3	7	954	46	0	1	12	12	0

ここには人生があります。
データをよく見ましょう。

この人労組に入ったけど
次の年に抜けてる、、、

この人、離婚したんだ、、、

ソース：Cameron and Trivedi (2010) *Microeconometrics Using Stata*, Stata Press.

- これは個人レベルのパネルデータの例：$N = 3$, $T = 7$ （このデータでは観測は一年おき）
- wageは賃金水準、wksは一年間で何週間働いたか。
- fem = 1（女性）fem = 0（男性）→　こういう変数を「**ダミー変数**」という。
- union = 1（労働組合に入っている）union = 0（入っていない）
- ms = 1（結婚している）ms = 0（していない）
- union も msも「**ダミー変数**」の例。
- edは学校に何年通ったか（学歴）。expは何年働いたか。

8

よく使う数学

- 回帰分析の理解のためには、数学の知識、確率の知識、統計の知識が必要になります。
- 以下では回帰分析でよく使われる数学を見ていきます。
  - 和のオペレーター
  - 線形関数
  - 二次関数
  - 対数関数
- 関数に関しては使うから重要というだけでなく、係数の読み方（分析結果の読み方）においても重要になります。

9

A.1　和のオペレーター

- **和のオペレター**は多くの数の和を含む式を簡潔に書くためのもの（それだけです）。
- 回帰分析で重要な役割を果たします。
  - 標本平均、標本分散、標本共分散などは、和のオペレーターを使って表記しないとごちゃごちゃになります。
- $x_1 + x_2 + \cdots + x_n$は和のオペレーターを使って次のように簡潔に表すことができます。

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

- 以下ではよく使うルールを見ていきます。

10

A.1　和のオペレーター

- **和のオペレターの特性１**：いかなる定数cについて

$$\sum_{i=1}^{n} c = nc$$

証明：

$$\sum_{i=1}^{n} c = \underbrace{(c + c + \cdots + c)}_{n\text{個}} = nc$$

11

A.1　和のオペレーター

- **和のオペレターの特性２**：いかなる定数cについて

$$\sum_{i=1}^{n} cx_i = c\sum_{i=1}^{n} x_i$$

- 証明：

$$\sum_{i=1}^{n} cx_i = (cx_1 + cx_2 + \cdots + cx_n) = c(x_1 + x_2 + \cdots + x_n) = c\sum_{i=1}^{n} x_i$$

- **和のオペレーターの特性３**：いかなる定数aとbについて

$$\sum_{i=1}^{n} (ax_i + by_i) = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i$$

12

- 注意１:「割合の和」は「和の割合」と等しくはない。すなわち

$$\sum_{i=1}^{n}(x_i / y_i) \neq \left(\sum_{i=1}^{n} x_i\right) \bigg/ \left(\sum_{i=1}^{n} y_i\right)$$

- これは簡単な例（n = 2）から分かる:

$$x_1 / y_1 + x_2 / y_2 \neq (x_1 + x_2)/(y_2 + y_2)$$

- 注意2:「２乗の和」は「和の２乗」と等しくはない。すなわち

$$\sum_{i=1}^{n} x_i^2 \neq \left(\sum_{i=1}^{n} x_i\right)^2$$

- これは簡単な例（n = 2）から分かる:

$$x_1^2 + x_2^2 \neq (x_1 + x_2)^2 = x_1^2 + 2x_1x_2 + x_2^2$$

A.1　和のオペレーター

- $x_1, \ldots, x_n$ が与えられたとき、それらの**平均**（average or mean）は:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- x_iがある変数（例えば教育の年数）の母集団からのサンプルの場合、これを**標本平均**（sample average or sample mean）という。

- 標本平均は**記述統計**の1つ。
  - 中心的な傾向をとらえる値

A.1　和のオペレーター

- **平均に関する重要な特性**:
  - それぞれの値 x_i から平均を引く（**平均からの偏差**）
  - それらの和は常にゼロになる。

$$\boxed{\sum_{i=1}^{n}(x_i - \bar{x}) = 0}$$

証明:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = \sum_{i=1}^{n} x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0$$

A.1　和のオペレーター

- 回帰分析で使われる「**（標本）平均からの偏差**」に関する**重要な特性**:は以下の通り:

$$\boxed{\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}$$

証明

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + (\bar{x})^2)$$

$$= \sum_{i=1}^{n} x_i^2 + 2\bar{x}\sum_{i=1}^{n} x_i + n(\bar{x})^2$$

$$= \sum_{i=1}^{n} x_i^2 - 2n(\bar{x})^2 + n(\bar{x})^2$$

$$= \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2$$

メモ：「標本分散」が含まれる式の展開のときに、このルールは役立ちます。

A.1　和のオペレーター

- 同様に以下のことが示せる:

$$\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})=\sum_{i=1}^{n}x_i(y_i-\overline{y})=\sum_{i=1}^{n}(x_i-\overline{x})y_i=\sum_{i=1}^{n}x_iy_i-n(\overline{x}\cdot\overline{y})$$

- メモ: 標本共分散を含む式の展開の時にこのルールは役立ちます。
- 繰り返し出てきますので、前頁のルールとともに「絵」として認識して覚えてしまうといいです。
- 証明: 復習問題

17

A.2　線形関数の特性

- ここでは係数の解釈の仕方について学びます。
  - ➢ 係数は符号だけでなく、大きさも大切です。
  - ➢ 係数が何を表しているかをきちんと理解することは、結果の解釈においてとても重要になります。
- xとyが変数で次のように関係しているとする:

$$y=\beta_0+\beta_1 x$$

- このときyはxの線形関数であるという。
- β_0とβ_1はこれらの変数の関係を表す係数(パラメータとも言う)。
  - – 前者は**定数項**（intercept）後者は**傾き**を表す係数（slope parameter）

18

A.2　線形関数の特性

- 線形関数の特徴は、yの変化（dy）はつねにβ_1にxの変化（dx）を掛けたもの。

$$dy=\beta_1 dx$$

- y のxに関する微分 dy/dx、これを回帰分析ではxのyへの限界効果と言います。
  - ➢ 解釈は「**xが1単位変化したときyがどれだけ変化するか**」
- 線形関数では、xのyへの限界効果は一定でβ_1 に等しくなります。

19

- 例：（線形家賃関数）家賃（C）と所得（I）との間に次のような関係があるとする（$表示）。

$$C=164+0.27\cdot I$$

- 所得(単位は$)の限界効果は0.27です。
- これをどう人に伝えるかですが、、、
- 「所得が1単位追加されると、家計は家賃に$0.27追加的に費やす」と説明する人、います。
- でも他人は「所得の1単位って？」と思います。
- ここの1単位はドルですから、「所得が1ドル追加されると、家計は家賃に27セント追加的に費やす」と係数を説明できるになること、これ大事です。

20

A.2 線形関数の特性

- 次は重回帰分析の係数の解釈の仕方です。
- 線形関数は2変数以上の場合にも定義することが可能。
- yが二つの変数x_1、x_2と次のように関係しているとする：

$$y = \beta_0 + \beta_1 x + \beta_2 x_2$$

- x_1とx_2がそれぞれdx_1とdx_2変化したとき、yの変化（dy）は

$$dy = \beta_1 dx_1 + \beta_2 dx_2$$

A.2 線形関数の特性

- x_2が一定の場合（$dx_2 = 0$）、$dy = \beta_1 dx_1$ となる。よってβ_1 はx_1についての傾きを表す。
- 従って、β_1は

「<u>x_2を一定にしたときに</u>、yがx_1によってどれだけ変化するか」

を表している。

- これを $\boldsymbol{x_1}$の偏効果（partial effect）ということもある。
- 「他の要因を一定にしたとき」の概念
- 偏微分を知っている人へ：これは単に偏微分です： $\partial y/\partial x_1 = \beta_1$
- β_2 も同様の解釈。$\partial y/\partial x_2 = \beta_2$

- 例：（CDの需要関数）CDの需要（D）、価格（P）、そして所得（I）との間に次のような関係があるとする（$表示）。

$$D = 120 - 9.8P + 0.03I$$

- ミクロ経済学で学習した需要曲線は、所得（またその他の要因）を一定としたときの、需要量と価格の関係。
- 需要曲線の傾きは–9.8: これは価格の需要量への偏効果。
  - <u>所得を一定とすると</u>、価格が$1増加したときにCDの需要は9.8枚減るということ。
- 所得の増加は、需要曲線を上側にシフトさせる。
  - 需要曲線の傾きは変化しない。

A.3 よく使われる関数

- 線形関数の特徴は、xのyへの限界効果が一定なこと。
- しかし現実は違うことあり。
  - 世の中の2つのことが何でもかんでも線形の関係にあるなんてことはない。
- 例えば、xが学習時間、yがテストのスコア。
- 最初の一時間の勉強がテストの点に与える効果と、百時間やった後の一時間の勉強がテストの点に与える効果、どちらが大きい？
  - まあ、普通は前者。
  - これはxが大きくなると限界効果が小さくなっていく例。
- <u>xのyへの限界効果がxの水準に依存する</u>ような状況をモデル化したい場合、回帰分析でどうするのか？
- 以下では、そういう時に使える関数（非線形関数）を紹介します。

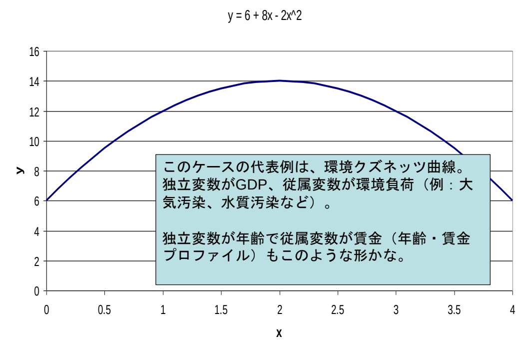
A.3 よく使われる関数：二次関数

限界効果が逓減するような状況をモデル化する一つの方法は、線形の関係に<u>2乗項をくわえる</u>こと（2次関数）。

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

- ちなみに「xの限界効果が逓減する」とは、「xが大きくなるにつれて、xの限界効果が小さくなっていく」ことのそれっぽい言い方。
- 以下の図はその一例。

25



26

A.3 よく使われる関数：二次関数

- $y = \beta_0 + \beta_1 x + \beta_2 x^2$ におけるxの<u>限界効果</u>は？
- これはyのxに関する微分：

$$dy / dx = \beta_1 + 2\beta_2 x$$

- $\beta_2 < 0$のとき、xが増加するとxの限界効果は減少していく。

27

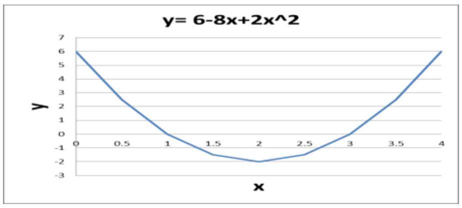
- 例：時間当たり賃金（W）と労働市場での経験年数（E）の間に次のような関係があるとする（$表示）。

$$W = 5.25 + 0.48E - 0.08E^2$$

- 1年目の労働経験は、賃金をおおよそ48セント上昇させる。
- それ以降の追加的な労働経験は賃金を増加させるが、その増分は減少していく。

28

- 従属変数と独立変数が逆U字の関係にある場合、2次関数を使うことでその関係を捉えることができる。
- また$\beta_2 > 0$の場合は、U字の関係を捉えることができる。



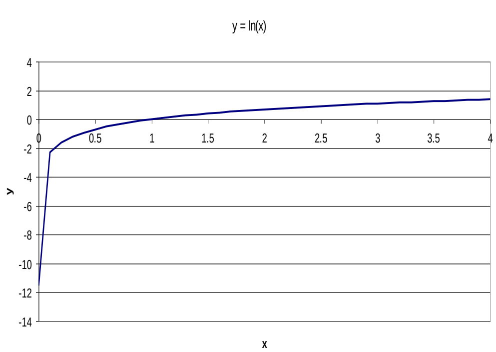
- この場合、限界効果は逓増する。

A.3　よく使われる関数：対数関数

- 自然対数は回帰分析で非常によく使われる。

$$y = \ln x$$

- これは$y = \log_e x$のこと。また単に$y = \log x$と書くこともある。回帰分析で使われる対数の底は常にe。
- 対数関数はx > 0についてのみ定義される。
- 重要：回帰分析では、正の値のみを取る連続変数をモデル化する場合、多くのケースで対数に変換したものを使う。



A.3　よく使われる関数：対数関数

- 対数関数の特性1：

$$\ln(x) < 0 \text{ for } 0 < x < 1$$
$$\ln(x) = 0 \text{ for } x = 1$$
$$\ln(x) > 0 \text{ for } x > 1$$

- ここら辺の特性は、本講義ではあまり使わないので、覚えなくてもいいです。

A.3　よく使われる関数：対数関数

- 対数関数の特性2：

$$\ln(x_1 x_2) = \ln x_1 + \ln x_2 \ \text{for } x_1 > 0, x_2 > 0$$
$$\ln(x_1 / x_2) = \ln x_1 - \ln x_2 \ \text{for } x_1 > 0, x_2 > 0$$
$$\ln(x^c) = c \ln x \ \text{for any } c$$

- この特性も、本講義ではあまり使わないので、覚えなくもていいです。

33

A.3　よく使われる関数：対数関数

- 対数は近似によく用いられる。本講義で重要なのは次の近似。
- $x_0 > 0$、$x_1 > 0$のとき

xはもともとx_0でx_1になった。
そのときのxの変化率。

$$\ln x_1 - \ln x_0 = \Delta \ln x \approx (x_1 - x_0) / x_0$$

- 「**対数の差は変化率**（の近似）」。
- 「**100掛ける対数の差は%表示した変化率**（の近似）」

34

A.3　よく使われる関数：対数関数

- この近似が良く使われる理由は？
- この理由を理解するために、「***x*に関する*y*の弾力性**」をまず定義する：

$$\frac{dy}{dx}\frac{x}{y}$$

- これは $= \dfrac{\frac{dy}{y}}{\frac{dx}{x}} = \dfrac{100\frac{dy}{y}}{100\frac{dx}{x}}$ 。この分子（分母）はy（x）のパーセント変化。
- だから**弾力性**とは

「 <u>xが1%変化したときにyが何%変化するか</u>」

- 弾力性は係数の解釈で使うので超重要。覚えること。

35

A.3　よく使われる関数：対数関数

- もしyがxの線形関数なら、弾力性は

$$\frac{dy}{dx}\frac{x}{y} = \beta_1 \frac{x}{y} = \beta_1 \frac{x}{\beta_0 + \beta_1 x}$$

となりxの値により変わる。

- 回帰分析では<u>弾力性が一定のモデル</u>がしばしば推定され、対数関数はそのような場合に使われる。
- 以下では、<u>弾力性が一定のモデル</u>がどのようなものか見てみよう。

36

A.3 よく使われる関数：対数関数

- xとyの変化率（%表示）の近似は、それぞれ100Δln$x$と100Δln$y$で与えられる。
- これよりxに関するyの弾力性は近似的にΔln$y$/Δln$x$ で表すことができる。
- 従って弾力性が一定（β_1）のモデルは

$$\ln y = \beta_0 + \beta_1 \ln x$$

の形をとる。

- この形の回帰モデルは非常によく推定される。
- 係数の解釈の仕方は、「xが1%変化したときyは β_1 %変化する」

37

- 例（価格弾力性が一定の需要関数）

$$\ln D = 4.7 - 1.25 \ln P$$

- この需要関数において、需要の価格弾力性は-1.25で一定。
- このモデル、係数を見たとき、

「1%の価格の上昇は需要を1.25%減少させる」

と読めるようになる、人に説明できるようになること。

38

A.3 よく使われる関数：対数関数

- 対数関数は実証分析において次の形のモデルもよく使われる。

$$\ln y = \beta_0 + \beta_1 x$$

- このときΔln$y$ = $\beta_1 dx$ → 100Δln$y$ = 100$\beta_1 dx$
- 左辺はyの変化率（%表示）なので、このモデルは

「xが一単位変化したときにyが100β_1%変化する」

ことを表している。

39

A.3 よく使われる関数：対数関数

- 例：（対数賃金関数）時間当たり賃金（W）と教育年数（EDU）の間に次のような関係があるとする。

$$\ln W = 2.78 + 0.094 EDU$$

- このとき教育のリターン（１年追加的に教育を受けると何%時間当たり賃金が増えるか）は9.4%で一定。

40

まとめ：係数の読み方

従属変数	独立変数	パラメータの解釈
Level	Level	Xが一単位増えるとYがβ変化する。
Log	Level	Xが一単位増えるとYが100β%変化する。
Log	Log	Xが1%増えるとYがβ%変化する（弾力性）。

注：Levelとは変数そのままということです。

連続変数はなぜ対数を取ることが多い？

- 「回帰分析では、正の値のみを取る連続変数をモデル化する場合、多くのケースで対数に変換したものを使う」
  - ➢ これは従属変数でも独立変数でもです。
  - ➢ 対数を取る（ことが多い）理由はいくつかあります。調べてみてください。
- ここでは教科書ではあまり説明されていないけれど、個人的には大切と思われるものについて説明します。
- log-logの場合について考えてみましょう。議論の簡単化のために定数項はとっちゃいましょう。

$$lny = \beta lnx$$

- 左辺は対数なので、このままではxと<u>yそのもの</u>の関係は分かりづらいですね。左辺がyになるように式をいじってみましょう。

連続変数はなぜ対数を取ることが多い？

- そのために次のルールを使います。

(1) $e^{lnA} = A$
(2) $clnx = lnx^c$

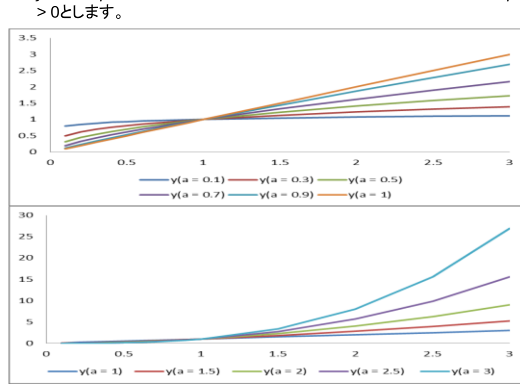
- $lny = \beta lnx$ だから$e^{lny} = e^{\beta lnx}$
- 左辺はルール（1）からy
- 右辺はルール（2）と（1）を使って、$e^{\beta lnx} = e^{lnx^\beta} = x^\beta$
- すなわち$lny = \beta lnx$ は

$$y = x^\beta$$

という関係を表していることになりますね。

- ➢ この式の両辺に対数を取ると$lny = \beta lnx$ になることを確認してください。

- $y = x^\beta$ のβの値をいろいろ変化させてグラフを書いてみます。ここでは$\beta > 0$とします。



- これらの図から、βの値によって、xがyに与える限界効果が逓減、一定（に近い）、また逓増したりすることが分かります。
- 従って、log-logは、level-levelに比べると、従属変数と独立変数の関係をより柔軟に捉えることができるということが言えると思います。
- 変数を対数変換して使うことが多い他の理由としては、解釈の容易さが挙げられる。
- logだと係数をパーセント（比率）で解釈できる。よって、係数の解釈が変数の単位に依存しない。
- 一方、levelの場合、変数の単位によって（例えばドルとセント）で係数の大きさが変わる。従って、係数の解釈が単位に依存することになる。

45

復習問題

- 横断面データ、時系列データ、パネルデータをそれぞれ言葉で説明しなさい。
- ダミー変数とは何か、言葉で説明しなさい。
- 以下の式が成り立つことを証明しなさい。

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}x_i(y_i-\bar{y})=\sum_{i=1}^{n}(x_i-\bar{x})y_i=\sum_{i=1}^{n}x_iy_i-n(\bar{x}\cdot\bar{y})$$

- 「xのyへの限界効果」とは何か、言葉で説明しなさい。

46

復習問題

- 二次関数のモデルがどのような時に使われるか説明しなさい。
- 弾力性とは何か、言葉で説明しなさい。
- 以下の表の__________を埋めなさい。

従属変数	独立変数	パラメータの解釈
Level	Level	Xが_____増えるとYが_____変化する。
Log	Level	Xが_____増えるとYが_____変化する。
Log	Log	Xが_____増えるとYが_____増える変化する。

47