回帰分析I

4. 記述統計
Stataの使い方イントロ

1

id	t	wage	wks	fem	union	ed	exp	ms
1	1	260	32	0	0	9	6	1
1	2	305	43	0	0	9	7	1
1	3	402	40	0	0	9	8	1
1	4	402	39	0	0	9	9	1
1	5	429	42	0	0	9	7	1
1	6	480	35	0	0	9	8	1
1	7				0	9	9	1
2	1				0	11	30	1
2	2				0	11	31	1
2	3				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					1
3	5	809	52					1
3	6	879	52	0	1	12	11	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.

- 前回、こんな感じでまずデータを味わうことをお勧めしました。
- しかし一つ一つ味わっていくのは大変です、とくにデータが大きいと。
- なので、データの特徴・傾向をつかむことが必要となります。

2

記述統計

- そのため、分析の第一歩として、記述統計を計算します。
- これは、標本平均、標本分散、最小値、最大値などで、データの示す特徴・傾向を知るためのものです。
  - 「データの特徴・傾向はあまり知らない、でも回帰分析はした」という人よくいます。
  - が、適切なモデリングのためには、データの特徴・傾向をよく知ることが必要になります。
  - 実証分析すればするほど記述統計の大切さに気が付いてきます。
- 以下では、Stataの使い方を学びながら、記述統計について見ていきましょう。

3

Stata はじめの一歩

- 教科書（松浦寿幸著「Stataによるデータ分析入門」東京図書）の第一章の内容です。
- データは通常EXCEL形式かカンマ区切り形式（CSV形式という）のいずれかで保存されていることが多いです。
- ここではCSVデータ（1994年と2004年の神奈川県藤沢市の家賃データ）

  rent-shonandai.csv

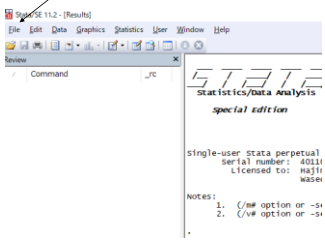
  をStataに読み込んでみましょう。
- その前に、 rent-shonandai.csvをよく見ましょう。
  - 一行目は変数名です。8変数あります。*N*=70です。
  - ブランク（数字が入っていないセル）があります =**欠損値**です（データが無い）。
- 読み込ませ方は少なくとも2種類ありますが、ここではメニュー・ウインドウを使う方法を紹介します。
  - もう一つの方法、insheetコマンドを使う方法、については教科書を参照してください。

4

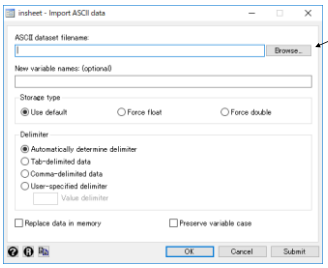
データ（csvファイル）の読み込ませ方

- File → Import → ASCIIdata created by a spreadsheet
- ASCIIdata data filename で　Browse … をクリック
- ファイル名（N）の右側にあるRaw Files（*, raw）をComma Separated Values （*, csv）に 変更
- データが置いてあるフォルダーに移動して、そのファイルをダブルクリック。
- 最後にDelimiter で Comma-delimited data を選択
- OK
- 以上です。

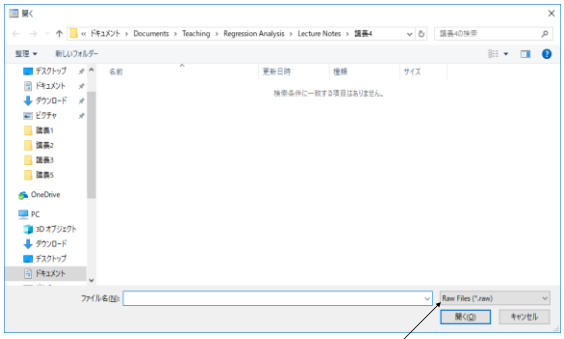
まずここクリック



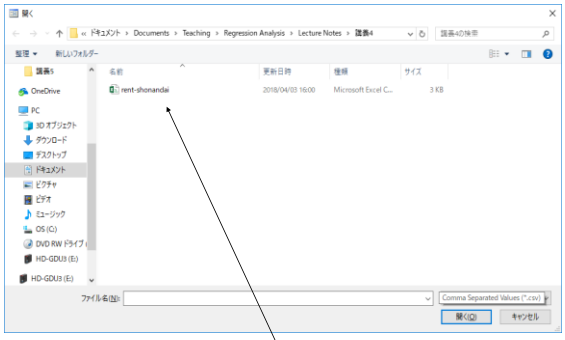
File → Import → ASCIIdata created by a spreadsheet



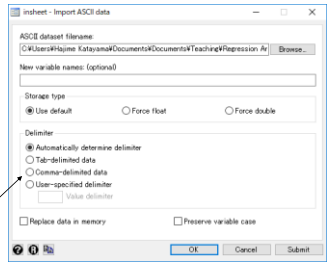
クリック



Comma Separated Values に変更



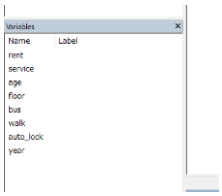
9



その後、OK。

10

画面左下：Variable ウィンドウ



これらがデータの中の変数

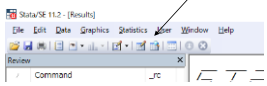
11

- ここで変数の定義を与えておきます。
  - rent: 賃貸料（単位：万円）
  - service: 管理費（単位：万円）
  - age: 築年数（単位：年）
  - floor: 占有面積（単位：m^2）
  - bus: 最寄り駅（湘南台）までのバス所要時間（単位：分）
  - walk: 徒歩分数（分）
  - auto_lock: オートロックの有無
  - year: 調査年
- ここでは、このデータセットを使った分析の目的を、「家賃はどのような要因によって決定されるのか？」とにでもしておきましょう。
- その中でも、特に、専有面積に興味があるとします。
  - これはこの講義ノートの中だけのことです。
  - 「家賃と占有面積の関係」が研究のトピックだとしたら、あまり面白くないトピックと言えます。
- 家賃は「賃貸料＋管理費」と定義することにします。
- 従って、この分析における従属変数は「賃貸料＋管理費」です。

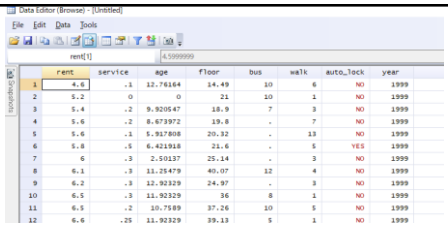
12

読み込んだデータの確認

- データをStataに読み込ませたら、まず最初にやることは、読み込んだデータの確認です。
  - これ、本当に大切です。
  - というのは、CSVファイルやEXCELファイルの方に何らかの問題があって、Stataがこちらの意図しない形でデータを読み込むことがあるからです。
  - 例えば、こちらは数字のつもりなのに、Stataはテキストとして認識したり、、、
  - 最後の行が読み込まれていなかったり、、、
  - それに気づかず分析して、見当違いの結果を得ることも。
- 確認の仕方ですが、、、



- Data Editor (Browse)
- この画面ではデータの加工はできない



	rent	service	age	floor	bus	walk	auto_lock	year
1	4.6	.1	12.76164	14.49	10	6	NO	1999
2	5.2	0	0	21	10	1	NO	1999
3	5.4	.2	9.920547	18.9	7	3	NO	1999
4	5.6	.2	8.673972	19.8	.	7	NO	1999
5	5.6	.1	5.917808	20.32	.	13	NO	1999
6	5.8	.5	6.421918	21.6	.	5	YES	1999
7	6	.3	2.50137	25.14	.	3	NO	1999
8	6.1	.3	11.25479	40.07	12	4	NO	1999
9	6.2	.3	12.92329	24.97	.	3	NO	1999
10	6.5	.3	11.92329	36	8	1	NO	1999
11	6.5	.2	10.7589	37.26	10	5	NO	1999
12	6.6	.25	11.92329	39.13	5	1	NO	1999

- CSVファイルと同じか確認しましょう。少なくとも最初の数行、最後の数行が同じかどうかは要確認。
- auto_lock 変数は赤字で表示されています。これは文字情報（ここではYESとNO）から構成される変数であるとStataが認識していることを示します。
- もともと数値が入っていないところ（欠損値）は、"."（ピリオド）になります。

宿題：

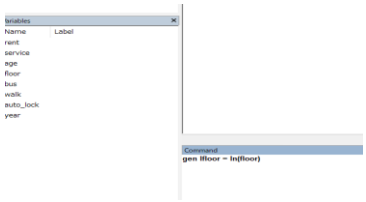
- 教科書（松浦寿幸著「Stataによるデータ分析入門」東京図書）の第一章を読む。
- とりあえず一通りやってみる。
- listコマンド（30ページ）を使ってみる。

「変数の置き換え」と「新しい変数の作成」

- これから、教科書（松浦寿幸著「Stataによるデータ分析入門」東京図書）の第二章の内容を扱います。
- 分析に際して、変数を置き換えたい、また既存の変数から新しい変数を作りたいことが頻繁にあります。
- まず「既存の変数から新しい変数を作りたい」ときによく使われるコマンドは generate です（省略してgenだけでもOKです）。
- 例えば、ここでは、floor（占有面積）の自然対数を新しい変数として作りたいとします。
- Commandウィンドウに以下のように書いてリターンして下さい。

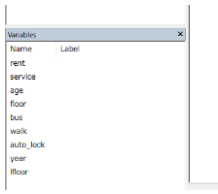
gen lfloor = ln(flootr)

generateコマンド



- 最初のgenはコマンドです。「以下のように新しい変数を作ってください」とStataにお願いします。
- 次のlfloorは新しく作りたい変数の名前（こちらが名付けます）で、それは（=）floor変数を対数変換したもの（ln(floor)）にして下さいと、お願いを具体化しています。

17



- 新しい変数ができました。
- Data Editor (Browse)を使って確認してみましょう。
- 同じ要領で、今度はfloor変数を二乗したものを作ってみましょう。新しい変数の名前はfloorsqにしましょう。
- gen floorsq = floor^2（gen floorsq = floor*floor でも同じものが作れます）。
- Data Editor (Browse)を使って確認しましょう。

18

- 今度は複数の既存の変数から新しい変数を作ってみましょう。
- 「占有面積当たりの賃料」を新しい変数として作ってみましょう。
- 新しい変数の名前は、rentperfloor とでもしましょうか。
- gen rentperfloor = rent/floor で作れます。
- こんな感じで新しい変数を作ることができます。
- ＋（足す）、－（引く）、*（かける）、/（割る）、^2（二乗）、ln(・)（自然対数）など使えます（他にもあります）。

19

replaceコマンド

- replaceコマンドも非常によく使われます。
- replaceコマンドは既存の変数を加工するコマンドです。
- 具体例を見ていきましょう。
- 入居者が負担する金額は、賃料（rent）と管理費（service）の合計です。
- この二つを合計したもの（ここでは便宜上、「家賃」と呼ぶことにします）に興味があるとします。
- 先ほどの要領でgenerateコマンドを使って、新しい変数（例えば、rent_service）を作ってもいいです。
- gen rent_service = rent + service ですね。

20

replaceコマンド

- ここでは、rentをrentとserviceを足したものに置き換えたいとします。
- その時は、

replace rent = rent + service

とCommandウィンドウに書いてリターンすればいいです。

- Data Editor (Browse)を使って確認してみましょう。
- 新しく変数は増えていませんね。
- ただし現在のrent変数は、以前のrent変数とは別物です。
- 現在のrent変数は、以前のrent変数とservice変数を足したものに置き換えられています。
  - CSVファイルを見て確認しましょう。

21

replaceコマンド

- もう一回replaceコマンドを使ってみます。
- いまデータ上では、バスを利用しない物件ではbusが欠損値（"."）になっています。
- この欠損値をゼロに置き換えたいとします。
- これは

replace bus = 0 if bus == .

でできます。バス変数をゼロにしてください（ replace bus = 0 ）、ただしbus変数が欠損値の時だけです（if bus == .）

とStataにお願いしていることになります。

22

if

- 先ほど、if ~ によって条件をつけました。
- 多くのコマンドは　if ~　と一緒に使えます。
- 「～の条件の時にだけ（コマンド）する」とStataに頼むことができます。
- if 以下は、

AとBが同じ	==	:	if A == B
AがBより大きい	>	:	if A > B
AがB以上	>=	:	if A >= B
AがBより小さい	<	:	if A < B
AがB以下	<	:	if A <= B

などがよく使われます。

- 複数の条件を組み合わせることもできます。&が「かつ」、|が「または」です。

23

gen、replace、ifの応用例（出現頻度：高）

- gen、replace、ifを使って新しい変数を作ることが良くあります。
- ここではcategoryという新しい変数を作りたい、その変数はrentが6万円以下なら1、6万より高いが9万円以下なら2、9万より高いなら3をとるもの、としましょう。
- いくつか作り方はありますが、一つの作り方は以下のものです。
- まず、gen category = .
- これでcategoryという変数ができました。値はすべて .（欠損値）です。
- 次に　.（欠損値）をreplaceコマンドとifを使って置き換えていきます。

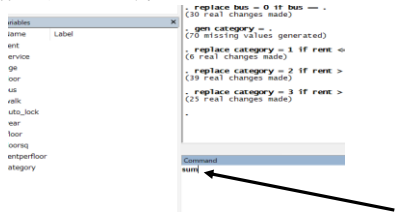
24

- replace category = 1 if rent <= 6
- これは「rent変数が6万以下なら、category変数を1にしてください。」とお願いしています。
- 次に、replace category = 2 if rent > 6 & rent <= 9
- これは「rent変数が6万より高くかつ9万以下なら、category変数を2にしてください。」とお願いしています。
- 最後に、 replace category = 3 if rent > 9
- これで完了です。
- Data Editor (Browse)を使って確認しましょう。
- より複雑な例（といっても少しだけですが）が教科書の40~41ページにあります。必ず見ておいてください。

25

SUMMARIZE

- 次に記述統計量の計算の仕方を紹介します。
- summarizeコマンドを使います。
- 画面の下の方にあるCommand ウィンドウにsum（ summarizeと書いてもOK）と書いてリターンします。



26

Variable	Obs	Mean	Std. Dev.	Min	Max
rent	70	8.716429	2.601055	4.7	18
service	70	.26	.2500145	0	.9
age	70	7.705988	8.264705	0	50
floor	70	47.53186	18.85208	14.49	86
bus	70	6.142857	5.866351	0	15
walk	70	4.514286	3.984546	1	18
auto_lock	0				
year	70	2001.571	2.517023	1999	2004
lfloor	70	3.769744	.4530517	2.673459	4.454347
floorsq	70	2609.601	1838.783	209.9601	7396
rentperfloor	70	.1936297	.0536246	.1245283	.32
category	70	2.271429	.611992	1	3

- Obsは観測値数。
- auto_lockはテキストデータとして認識しているので、観測値数がゼロになっている。
- この変数を数値データに変換する方法は、教科書の39ページ参照。
- Meanは標本平均、Std.Dev.は標本標準偏差、Minは最小値、Maxは最大値。

27

標本平均

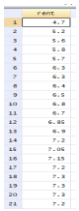
- rent（家賃プラス管理費）の標本平均は8.72万円。
- 計算方法？
- ここではー般化してN人からなる横断面データを想定。
- ある変数Xについて考える。
- ID=1の人の変数Xの値をX_1、ID=2の人の変数Xの値をX_2、、、、ID=Nの人の変数Xの値をX_Nと表しましょう。
- データは手短に書けば、X_i（$i = 1,2, \ldots, N$)です。
- 変 数 Xの標本平均は、一般的には $\bar{X}$ （エックス・バーと読むこともあります）と表記し、

28

標本平均

$$\bar{X} = \frac{1}{N}(X_1 + \cdots + X_N) = \frac{1}{N}\sum_{i=1}^{N} X_i$$

- 標本平均はその変数の位置の尺度で、代表的・典型的な値を測るものです。



- 今の例では、下まで足していって70で割ったもの。

29

標本分散

- 標本分散（s_X^2）は次の式で定義されます。

$$s_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

- 標本分散は散らばりの尺度の一つで、X_i（i = 1,2, …, N）がどれだけバラついているかを測るものです。
- まずそれぞれのX_iから標本平均$\bar{X}$が引かれる。
  - （標本）平均からどのくらいずれているか
- そしてそれぞれを二乗。
  - 二乗するのは正のズレも負のズレも等しく正の値で評価するため。
  - 二乗したものはズレのある種の測度。
- そしてそのズレの測度の"平均"を計算。

問：二乗しないでそのまま足すとどうなる？

注：NではなくN-1で割っているのにはちょっとした理由あり。今は気にしないこと。

30

標本標準偏差

Variable	Obs	Mean	Std. Dev.	Min	Max
rent	70	8.716429	2.601055	4.7	18
service	70	.26	.2500145	0	.9
age	70	7.705988	8.264705	0	50
floor	70	47.53186	18.85208	14.49	86
bus	70	6.142857	5.866351	0	15
walk	70	4.514286	3.984546	1	18
auto_lock	0				
year	70	2001.571	2.517023	1999	2004
1floor	70	3.769744	.4530517	2.673459	4.454347
floorsq	70	2609.601	1838.783	209.9601	7396
rentperfloor	70	.1936297	.0536246	.1245283	.32
category	70	2.271429	.611992	1	3

- レントの標本標準偏差は2.6万円。
- 標本標準偏差も散らばりの尺度。
- 一般的に、標本標準偏差は標本分散の正の平方根、すなわち$\sqrt{s_X^2}$

31

標本標準偏差

- 分散に平方根を取る理由は？
- 標本分散は計算の際に二乗を伴うために、その単位が変数単位の二乗になってしまう。
- 例えば、Xが「万円」単位で測られているとすると s_X^2 の単位は「万円^2」に。
- 標準偏差は平方根を取っているから、もとの測定単位と同じになる。

32

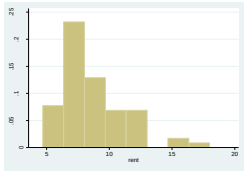
最小値、最大値

- 最小値、最大値の求め方については特に説明することなし。
- 最小値、最大値の値を知ること、それ自体も大切なことだが、他にそれらの値を知りたい理由があります。
- それはデータの誤入力の可能性を知れること。
- 月の家賃データで、最小値が20円だとしたら、それはかなり変。
- また最大値が3500万円だとしたら、それも変（場所は藤沢）。
- そういう場合は誤入力の可能性を疑うこと。
  - 調べてみる必要あり。
  - 場合によっては、分析から落とすことも。

33

ヒストグラム

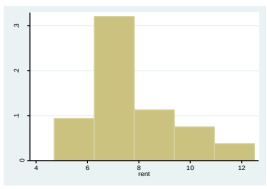
- hist rent　でrentのヒストグラムが書けます。
- この縦軸は割合です（頻度にすることもできます）。



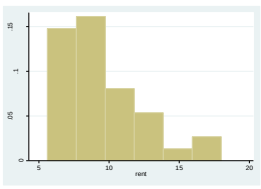
- 回帰分析をする前に、特に従属変数のヒストグラムを見ることは重要。
  - データの散らばり具合を視覚的に把握する。
  - データの中心がどれぐらいの位置にあるか視覚的に把握する。
  - 異常値（外れ値）の存在を視覚的に確認する。

34

- この家賃データは1999年と2004年のものなので、それぞれの年で分けてヒストグラムを書いてみましょう。
- ifを使ってみましょうか。
- hist rent if year == 1999
- hist rent if year == 2004



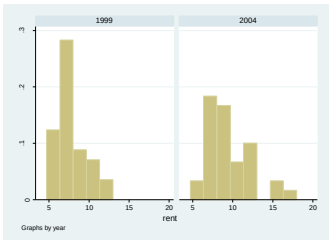
1999年



2004年

35

- hist rent, by(year) とやると年ごとに分けてヒストグラムを書いてくれます。



- 1999年と2004年では、家賃の分布の形がかなり違うのが分かります。

36

- ついでに、rentの記述統計も年ごとに出してみましょう。
- ここではおさらいの意味も込めてifを使って見ましょう。

```
sum rent, if year == 1999
sum rent, if year == 2004
```

  - ちなみにsumの後に変数を指定すると、Statはその変数だけの記述統計を与えます。
  - 特に指定しないと（先ほどは指定しませんでした）、データセットのすべての変数の記述統計が計算されます。

```
. sum rent if year == 1999
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        rent |         34    7.727941    1.878636         4.7       12.5

. sum rent if year == 2004
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        rent |         36        9.65    2.857346         5.6         18
```

- この記述統計から何が言えますか？

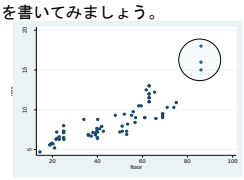
二変数の関係：散布図

- 回帰分析をする前には、各変数の平均、分散、最大値、最小値などをまず見てみること。
- では、その次は？
- 視覚的に変数間の関係性を把握することが大切。
- そのためには**散布図**を書く。
- 変数のコンビネーションすべてについて散布図を書く必要は無いです（もちろん書いても構いませんが）。
- 従属変数と重要な（分析において特に興味の対象である）説明変数の散布図を書けば十分です。

二変数の関係：散布図

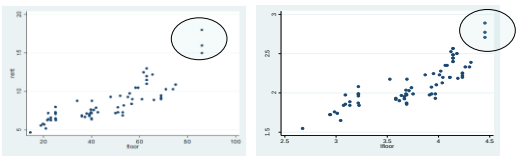
- それでは書いてみましょう。
- 特に興味がある説明変数は占有面積だとしました。
- なので、家賃と占有面積の散布図を書いてみましょう。
- twoway (scatter rent floor)

で書けます。



- 予想通りですかね。
- ただし、個人的には、これが少しだけ気になりますね

- 二変数ともに連続変数なので、それぞれを対数にした場合の散布図も書いてみましょう。
- gen lrent = ln(rent)
- floor変数の方は、すでに対数作ってありますね。
- twoway (scatter lrent lfloor)



- 3つの観測値の「外れ値」感、対数を取った場合には少し弱まりますね。

標本共分散

- 変数間の統計的な関係を数値的に表す指標の一つは**標本共分散**。
- 二つの変数をX_iとY_iと置けば、標本共分散（s_{XY}）は以下のように定義される。

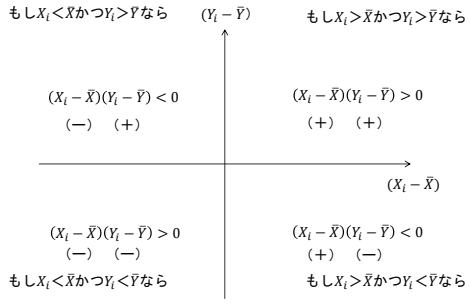
$$s_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$$

- 何を測っている？
- 変数のペア（X_i, Y_i）を考えよう。
- それぞれからそれぞれの標本平均を引いたものが、

$$(X_i - \bar{X})(Y_i - \bar{Y})$$

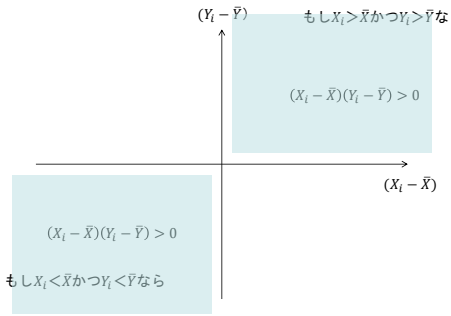
41

標本共分散



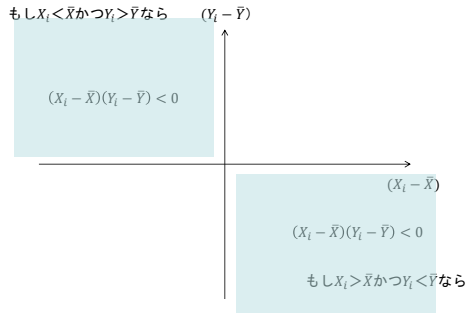
42

ざっくり言って、多くのデータがここにあれば、すなわち
$(X_i - \bar{X})(Y_i - \bar{Y}) > 0$ であるなら、XとYは右上がりの関係にあるということ。



43

ざっくり言って、多くのデータがここにあれば、すなわち
$(X_i - \bar{X})(Y_i - \bar{Y}) < 0$ であるなら、XとYは右下がりの関係にあるということ。



44

$s_{XY} > 0$ならXとYは**正に相関（右上がりの関係）**しているという。

$s_{XY} < 0$ならXとYは**負に相関（右下がりの関係）**しているという。

- 注意：s_{XY}は変数XとYの測定単位に依存する。従って、理論上の上限、下限は不明。
- そのため、共分散から相関の正負は分かっても、相関の強弱は分からない。

45

標本相関係数

- 標本共分散は測定単位に依存するため、相関の大小については言うことができない。
- 相関の大小について言うことを可能にする指標は、**標本相関係数**（r_{XY}）。

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{X\text{と}Y\text{の共分散}}{X\text{の標準偏差} \times Y\text{の標準偏差}}$$

- この指標は上限・下限あり：$-1 \leq r_{XY} \leq 1$

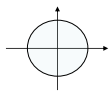
-1に近いほど強い負の相関、 1に近いほど強い正の相関

46

標本相関係数

- $r_{XY} = -1$なら、負の傾きの直線の上に、 (X_i, Y_i) $i = 1,\ldots,N$ がすべてのっている、ということ。
- $r_{XY} = 1$なら、正の傾きの直線の上に、 (X_i, Y_i) $i = 1,\ldots,N$ がすべてのっている、ということ。

注意：2変数の間に「相関が無い」からといって、その2変数が関係していないというわけではない。例えば、



この輪の上に等間隔に (X_i, Y_i) が並んでいたら、標本相関係数はゼロ。でも二つは強く関係してると思います（なぜ？）。

47

標本相関係数

- 分析の際、散布図を見た後は、標本相関係数を見ておく。
- Stataで標本相関係数を計算してみましょう。
- corr rent floor

	rent	floor
rent	1.0000	
floor	0.8454	1.0000

- 家賃と床面積の間には結構強い正の相関あり。
- corrの後に2変数以上書くこともできる。
- corr rent floor age

48

	rent	floor	age
rent	1.0000		
floor	0.8454	1.0000	
age	-0.3451	-0.0476	1.0000

「相関行列」と呼びます

- 家賃と床面積の間には強い正の相関有り（0.85）。
- 家賃と築年数の間には負の相関有り（−0.35）。
- 築年数と床面積の間には相関は無いかあっても弱い負の相関（−0.05）。
- 回帰分析の前に、従属変数と説明変数の相関を見て、ざっくり関係を知っておくことは大事。
- また説明変数間の相関もチェックしたい強い理由あり。
  - 重回帰分析の講義の時に、その理由について説明します。

49

宿題

- 教科書第2章の2.1-2.5を一通り読んでやってみること。

注：将来自分でStataを使って実証分析をするなら、

2.6「Do-fileによる作業のプログラム化」
2.7「log ファイルによる作業結果の保存」

の仕方を知っておくことは大切です。ただし本講義の試験的には全く関係ありません。

50