Universidade de São Paulo
Faculdade de Saúde Pública
Departamento de Epidemiologia

Programa de Verão, 2017

Stata Básico
Versões 9/10

Denise Pimentel Bergamaschi
José Maria Pacheco de Souza

2017

Índice

1. Iniciando o trabalho no *Stata*

Stata [Estata ou Esteita] – *StataCorp LP. 2007*

- *Intercooled Stata*

Existem versões do programa para três sistemas: *Windows, Unix* e *Macintosh*. Atualmente está na versão 14.

Este curso: Intercooled Stata versões 9/10 para sistema *Windows*.

Informações sobre o *Stata*, bem como atualizações, realização de cursos via *Internet*, livros, fórum de discussão, lista das dúvidas mais frequentes podem ser encontrados no *site* http://www.stata.com.

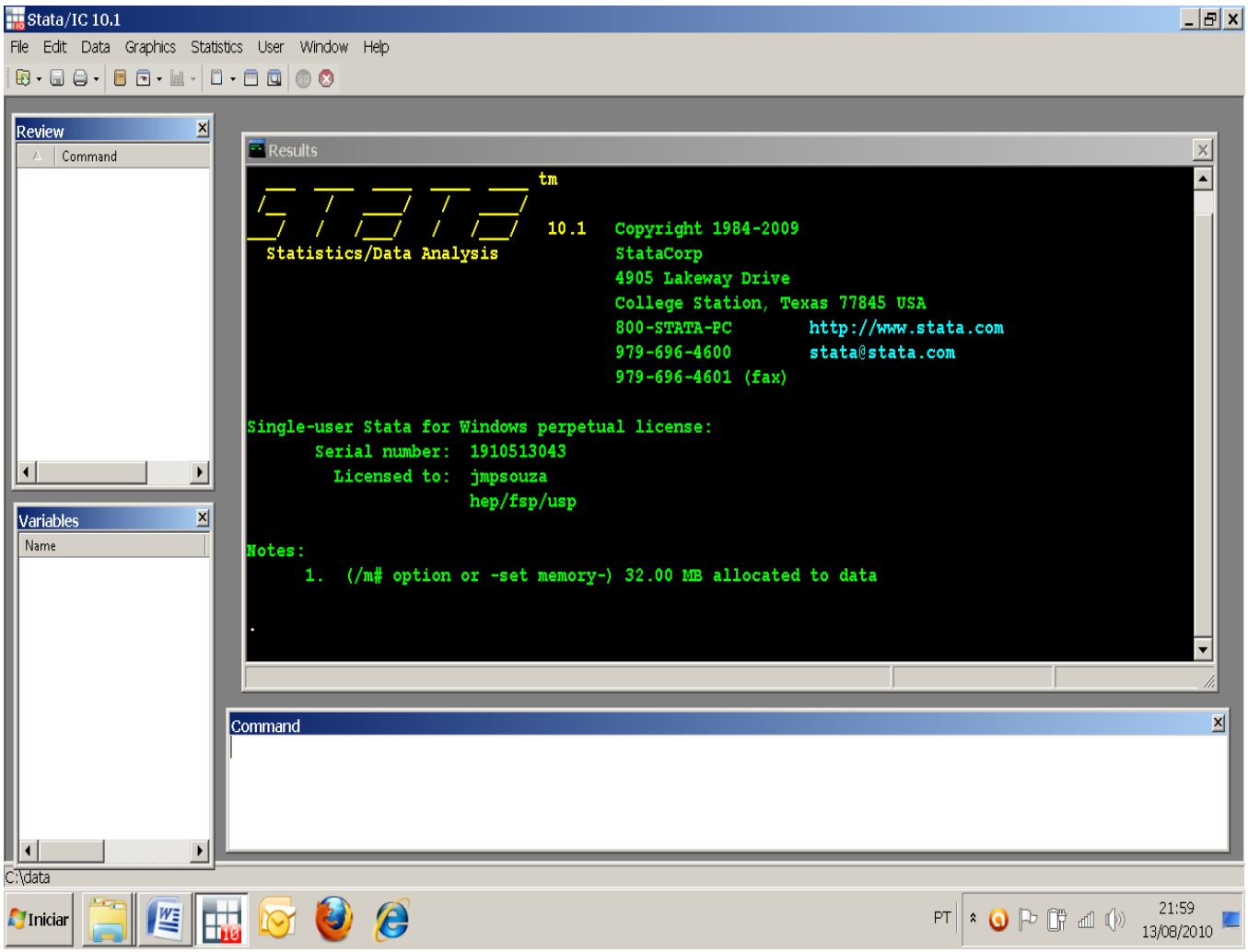
Os comandos podem ser dados via linha de comando ou via menu. O curso abordará comandos úteis em análise epidemiológica, com ênfase na linha de comando, mas também usando alguns menus.

Os arquivos de dados utilizados no curso e o manual básico estão na página

www.fsp.usp.br/~jmpsouza/Statabasico

1.1 - Iniciando o Stata

Abrir o programa. Obtém-se uma tela contendo quatro janelas, mais uma barra de menus e uma de ícones.



Atualização

É recomendável fazê-la de tempos em tempos. Utilizá-se o menu

Help →Official Updates (ou Check for updates)→ http://www.stata.com , seguindo as instruções. Serão atualizados o arquivo .exe e os arquivos .ado.

A finalidade de cada janela é apresentada a seguir:

Título	**Finalidade**
Review	Armazenamento dos comandos já utilizados. O comando pode ser reutilizado e corrigido utilizando-se o mouse ou as teclas **PgUp** (*page up*) e **PgDn** (*page down*)
Variables	Apresentação das variáveis no banco de dados.
Stata Results (fundo preto)	Apresentação dos resultados obtidos com a execução dos comandos
Stata Command	Para digitação dos comandos a serem executados. Digitar quando o *prompt* estiver ativo. Executar clicando a tecla **Enter.** Os comandos são em letras minúsculas.

1.2- Tipos de arquivos

O *Stata* trabalha com vários tipos de arquivos:

Tipo de arquivo	Extensão
Arquivo principal, executável	.exe
Arquivo que contém os dados	.dta
Arquivo que guarda os comandos e resultados obtidos durante a sessão de trabalho	.log; .smcl
Arquivo que contém comandos	.do
Arquivo que contém sub-rotinas	.ado
Arquivo de gráfico	.gph
Arquivo com gravação de construção de gráfico	.grec

1.3 – Sintaxe dos comandos

Os comandos seguem a forma:

[bysort varlist:] command [varlist] [weight] [if exp] [in range] [using filename] [,options]

onde

[bysort varlist:] instrui *Stata* para repetir o comando para cada combinação de valores nas variáveis listadas em *varlist*, que são ordenadas pelo sufixo "sort".

command é o nome do comando, ex: **list**

[varlist] é a lista de variáveis para as quais o comando é executado

[weight] permite que pesos sejam associados às observações

[if exp] restringe o comando a um subconjunto de observações que satisfazem a expressão lógica definida em *exp*

[in range] restringe o comando àquelas observações cujos índices pertencem a um determinado subconjunto

[using filename] especifica o arquivo que deve ser utilizado

[,options] são opções específicas de cada comando.

Ex: usando um banco de dados contendo as variáveis **x** e **y**
o comando para listá-las é: **list x y**

pode ser definida uma condição (if)**: list x y if x>y**

O programa diferencia entre letra maiúscula e minúscula. Todos os comandos e componentes são em letra minúscula.

A utilização do **Help** é fortemente recomendada; clicando-se em **Help** no menu principal, pode-se pesquisar qualquer comando utilizando-se a opção **Contents (todo o manual)**, **Search (palavras chaves)** ou **Stata command (comando)**.
O comando **findit <algo>** dá informações na internet sobre o item "algo".

O Stata funciona com o mínimo de memória e tamanho de matriz para tornar o programa mais ágil. Quando, ao abrir um banco de dados, surgir o aviso de memória insuficiente, a memória pode ser expandida pelo comando:

- **set memory 32m** (ou 16m, ou 64m etc) -A partir da versão 12, a alocação de memória é automática-

Quando, ao abrir um banco de dados, surgir o aviso que o tamanho da matriz é insuficiente, a matriz pode ser aumentada pelo comando:

- **set matsize 400** (ou 200, ou outro tamanho)

Os comandos que iniciam com **set** alteram o *default* de configurações do programa e devem ser realizados sem arquivo aberto.

1.4 - Abrir um arquivo log

Logo que for iniciado o trabalho no *Stata*, é aconselhável abrir um arquivo **log,** que armazenará todos os comandos e seus resultados (com exceção de gráficos). Gráficos podem ser copiados em arquivo doc.

Ou digitar

- log using <diretório:\nome do arquivo>

O arquivo **log** é um arquivo de tipo somente texto e não permite alteração. Caso seja de interesse, pode-se abri-lo em um editor de textos (*Word for Windows, bloco de notas*) e salvá-lo com extensão .doc ou .txt para ser manipulado segundo a necessidade.

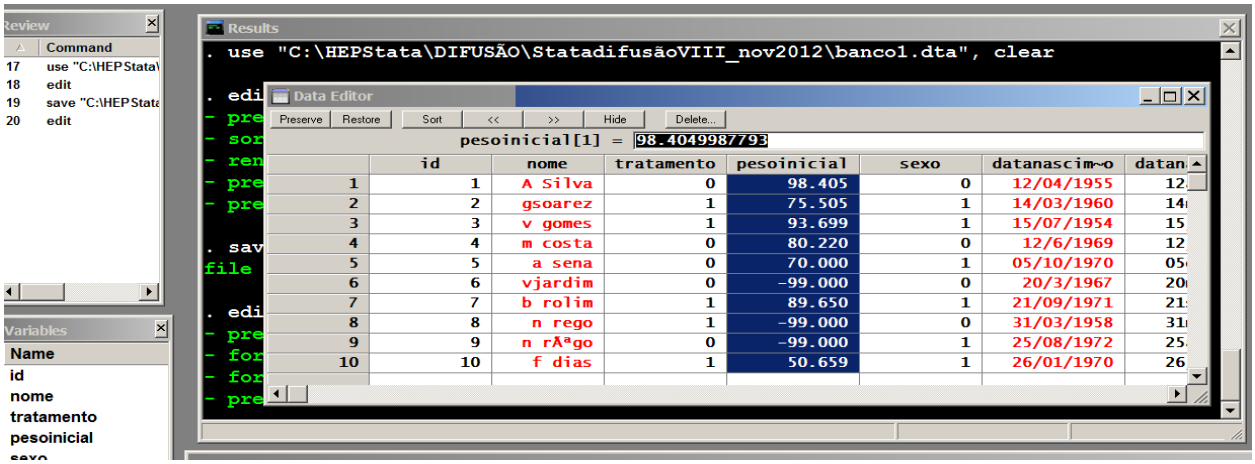
1.5 – Criando um banco de dados no *Stata*

Criar um banco de dados na planilha do **Edit,** com as variáveis id, nome, tratamento, pesoinicial, sexo e datanascimento; dez pacientes, a seguir:

id	nome	tratamento	pesoinicial	sexo	datanascimento
1	A Silva	0	98.405	1	12/04/1955
2	G Soares	1	75.505	2	14/03/1960
3	V Gomes	1	93.699	2	15/07/1954
4	M Costa	0	80.22	1	12/06/1969
5	A Sena	0	70.0	2	5/10/1970
6	V Jardim	0	-99	1	20/03/1967
7	B Rolim	1	89.65	2	21/09/1971
8	N Rego	1	-99	1	31/03/1958
9	F Dias	0	-99	2	25/08/1972
10	H Bastos	1	50.659	2	26/01/1970

-99 = sem informação

Abrir modo de edição clicando sobre o ícone **Data editor** (9º ícone do menu com desenho de uma planilha) e digitar os dados de cada registro. Usar **Tab** para entrada horizontal e **Enter** para entrada vertical. Quando terminar, pressionar **Preserve** seguido de **Close** no menu do editor (ou pressionar o **X** do lado direito da tela).



O **Browser** (ícone do menu com desenho de uma planilha com uma lupa) tem como finalidade visualizar o banco de dados sem a modificação dos dados.

Também há a possibilidade de digitar os comandos **edit** ou **browse** para ter acesso às planilhas ou pelo menu **Data**. Digitando os comandos, podem-se selecionar as variáveis desejadas para aparecer na planilha na linha de comando.

- **edit <var>**
- **browse <var>**

1.6 – Salvamento e leitura de banco de dados

Para salvar um arquivo novo (3 maneiras diferentes):

O arquivo deve ser salvo utilizando a caixa de diálogo do menu File, na seqüência:

- File, **Save as**, diretório - **cursostata**, nome do arquivo: **banco1**

- Ou pressionando o ícone disquete

- Ou digitando na linha de comando:

- **save c:\cursostata\banco1.dta**

Para fechar um arquivo sem salvar e sem abrir outro arquivo:

- **clear**

- Ou abrir outro arquivo, sem salvar o atual:

- **use c:/cursostata/outrobanco.dta, clear**

Para abrir um arquivo

- Pressionar o *mouse* sobre **File** seguido de **Open**. Seleciona-se o diretório que contém o nomedearquivo ***.dta***, marca-se o arquivo e seleciona-se **Open**.

- Ou pressionar o primeiro ícone do menu (arquivo amarelo)

- Ou digitar na linha de comando:

- **use c:\cursostata\nomedearquivo.dta, clear**

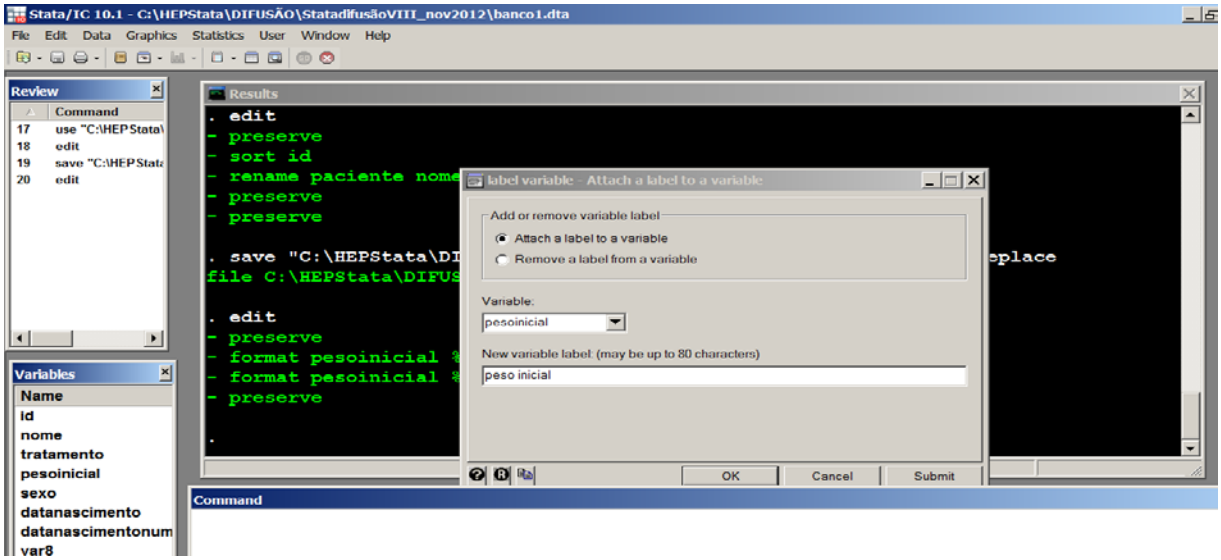
1.7 – Manipulação de variáveis

Há dois tipos de variáveis no Stata: string (caracteres, letras) e numérica. Estas variáveis são armazenadas de formas diferentes que requerem tamanhos diferentes nos registros de memória: *byte*, *int*, *long* e *float* para variáveis numéricas e *str1* até *str80* para variáveis string de tamanhos diferentes. Cada variável pode ter um nome associado a ela (rótulo, *label*) e ter um formato de apresentação.

Definir rótulo da variável:

- **label var pesoinicial "peso inicial"**

Ou Data → Labels → Label variables → Attach a label to a variable


Stata/IC 10.1 - C:\HEPStata\DIFUSÃO\StatadifusãoVIII_nov2012\banco1.dta
File Edit Data Graphics Statistics User Window Help
Review
Command
17 use "C:\HEPStata\
18 edit
19 save "C:\HEPStata
20 edit
Variables
Name
id
nome
tratamento
pesoinicial
sexo
datanascimento
datanascimentonum
var8
Results
. edit
- preserve
- sort id
- rename paciente nome
- preserve
- preserve
. save "C:\HEPStata\DI
file C:\HEPStata\DIFUS
. edit
- preserve
- format pesoinicial %
- format pesoinicial %
- preserve
.
label variable - Attach a label to a variable
Add or remove variable label
Attach a label to a variable
Remove a label from a variable
Variable:
pesoinicial
New variable label: (may be up to 80 characters)
peso inicial
OK
Cancel
Submit
Command

Descrever o formato e os rótulos das variáveis:

- **describe**

Ou Data → Describe data → *Describe data in memory*

Describe data in file

Describe data contents (codebook)

Obter resumo quantitativo dos dados:

- **sum**
- **sum, detail**

Comprimir o armazenamento, economizando memória:

- **compress**

Colocar as variáveis em ordem alfabética:

- **aorder**

Ordenar variáveis em ordem crescente, aninhando uma na outra:

- **sort** *varlist*

Ordenar variáveis em ordem crescente ou decrescente:

- **gsort [+|-]***varname* **[+|-]***varname* …

Listar os valores das variáveis:

- **list** *varlist*

Ou Data → List

- Pode-se selecionar variáveis. Se nenhuma variável for listada todas as variáveis serão listadas.

- As opções by|if|in podem ser usadas:

by – listar a variável y por categoria da variável x

if - listar a variável y se a condição for aceita

in – listar a variável y se a condição estiver no intervalo da variável x

- **sort tratamento**
- **list sexo if pesoinicial >=50 &tratamento==0**

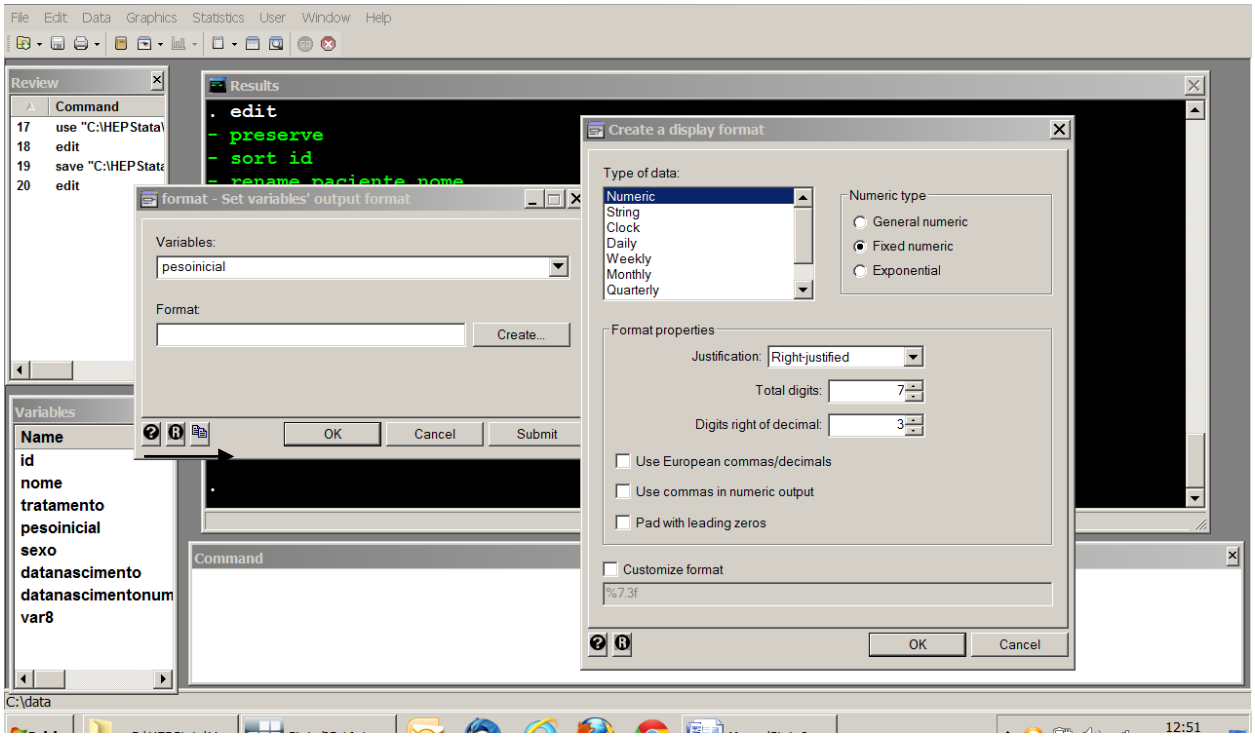
- **list sexo if pesoinicial>=50 &tratamento==1**
- **bysort tratamento: list sexo if pesoinicial >=50**
- **list sexo in 1/5**

Modificar formato de variável:

- **format pesoinicial %7.2f**

Ou

Data → Variables utilities → Set variables´ output format



Modifica o tamanho da variável numérica *pesoinicial*: 7 espaços antes da virgula e 3 casas decimais após a virgula em um formato fixo (f).

- **format nome %15s**

Modifica o tamanho da variável string *nome*: 15 espaços ao invés de 10.

Arredondando valores, usando o comando display e a função round():

- **display round(32.56004)** fica 33
- **display round(32.56004,.1)** fica 32.6

- **display round(32.56004,.01)** fica 32.56
- **display round(32.56004,.001)** fica 32.56
- **display round(32.56004,.0001)** fica 32.56
- **display round(32.56004,.00001)** fica 32.56004

Usando o comando generate, gen:

- **gen novavar=round(velhavar,.0001)**
- **gen novavar=round(velhavar)**

Valor inteiro

- **display int(32.56004)** fica 32
- **gen novavar=int(velhavar)**

Usando a função real()

- **gen varnumérica=real(varstring)**

Renomear variáveis

- **rename nome paciente**

Ou Data → Variables utilities → Rename variables

Variáveis numéricas

Valores faltantes (*missing*) são representados por pontos e são interpretados como valores muito grandes.

Substituir determinado valor por missing (.)

- **mvdecode pesoinicial, mv(-99)**

"-99" muda para "**.**"

Ou

Data → Create or change variables → Other variable transformation commands → Change numeric values to missing

Substituir missing (.) por determinado valor

- **mvencode pesoinicial, mv(-99)**

"." muda para "-99"

Ou

Data → Create or change variables → Other variable transformation commands → Change missing to numeric values

Recodificação de variáveis:

- **recode sexo 1=0 2=1**

Ou

Data → Create or change variables →Other variable transformation commands → Recode categorical variable

Definição de <u>rótulos</u> para categorias de variáveis:

- **label define rotulosexo 0 "masculino" 1 "feminino"**
- **label values sexo rotulosexo**
- **tab sexo**
- **tab sexo, nolabel**

Ou

Data → Labels & notes →**Define or modify Values labels** → Define value label

Data → Labels & notes → Assign value label to variable

OBS: Quando o rótulo for igual para várias variáveis é possível direcionar um único rótulo para todas estas.

Eliminar rótulos

- **label drop {lblname [lblname ...] | _all}**

Variáveis *string*

Variáveis *string* são utilizadas para variáveis com categorias não numéricas, sob a forma de palavras, ou, genericamente, um conjunto de caracteres, com ou sem sentido de palavra. São representadas por %# (# = nº de caracteres)

1.8 - Expressões

Existem expressões lógicas e algébricas, no *Stata*.

Expressões lógicas atribuem 1 (verdadeiro) ou 0 (falso) e utilizam os operadores:

Operador	Significado
<	menor que
<=	menor ou igual a
>	maior que
>=	maior ou igual a
= =	igual a
~= !=	diferente de
~	Não
&	E
|	Ou

Ex: **if (y~=2 & z>x) | x= =1**

Significa: se (**y** for diferente de 2 e **z** maior do que **x**) ou **x** for igual a 1

Expressões algébricas utilizam os operadores:

Operador	Significado
+ -	soma, subtração
* /	multiplicação, divisão
^	elevado à potência
sqrt( )	função raiz quadrada
exp( )	função exponencial

log10( )	função logarítmica (base 10)
ln( )	função logarítmica (base *e*) - logaritmo natural
logA(x)	(log10(x))/(log10(A))

1.9 - Observações índice e conjunto de valores

Observações índice

Cada observação está associada a um índice. Por exemplo, o terceiro valor da variável *x* pode ser especificado como **x[3]**. O macro **_n** assume os valores 1, 2, ..., _n para cada observação ordenada da variável x e **_N** é igual ao número total de observações. Pode-se referir à penúltima observação da variável *x* escrevendo-se **x[_N-1]**.

Uma variável indexada deve ficar do lado direito de uma asserção. Por exemplo, para substituir a terceira observação da variável *x* pelo valor 2 escreve-se:

- **replace nome=”joão” if _n==4**
- **replace nome=”vjardim” if _n==_N-6**

Ou

Data → Create or change variable → Change contents of variable

1.10- Variáveis data (calendário)

O *Stata* lê variáveis data como tempo decorrido (*elapsed dates)* ou **%d**, que é o número de dias contados a partir de 01 de janeiro de 1960. Assim,

0 corresponde a	01jan1960
1 corresponde a	02jan1960
.	.

. .
. .
15000 corresponde a 25jan2001

O *Stata* possui funções para converter datas em **%d**, para imprimir **%d** em formatos compreensíveis.

Variáveis datas devem ser definidas como variáveis *string* e depois convertidas para **%d**.

- **gen datanascimentonum =date(datanascimento,"DMY")**
  "dmy", nas versões <10
- **list datanascimento datanascimentonum**
- **desc**
- **format datanascimentonum %d**
- **list datanascimento datanascimentonum**
- **gen idade=(date("28/06/2004", "DMY")-datanascimentonum)/365.25**
- **list datanascimento datanascimentonum idade**

Ou

Data → Create or change variable → Create new variable

Data → Variables utilities → Set variables display format

- **gen datanum=date(datastring, "DMY#")**
- **gen datastring=string(datanum, "%td")**

2. Manipulação de dados

Abrir um arquivo c:\cursostata\sistolicainic.log para armazenar os resultados

Abrir o arquivo c:\cursostata\sistolicainic.dta

Os dados que serão utilizados nesta sessão constituem uma amostra de 58 pacientes hipertensos, do sexo feminino, que foram avaliados por 6 meses. As variáveis estudadas foram:

- **droga**: tipo de medicamento utilizado no período (1=nenhum; 2=tipo A; 3=Tipo B; 4=Tipo C)
- **sistolica**: incremento da pressão sistólica
- **idade**: idade em anos
- **salario**: renda do paciente (R$)
- **familia**: número da família (tem pacientes da mesma família).
- **pesoin:** peso inicial (kg) do paciente
- **pesointer:** peso (kg) do paciente após 3 meses de tratamento
- **pesof:** peso (kg) após 6 meses de tratamento

Para repetir comandos para variáveis ou categorias de variáveis, utilizar **by varlist;** os dados precisam estar ordenados antes disto, o que é feito utilizando o comando **sort**.

- **sort droga**

- **by droga: list sistolica**
- **bysort droga: list sistolica**

Ou

Data → Describe data → List data

* Ao comandar a listagem da variável sistólica por tipo de droga pelo menu, a variável droga será automaticamente ordenada antes, não necessitando o comando *sort*.

2.2 – Gerando, criando variáveis

O comando **generate** cria uma nova variável igualando a uma expressão que é construída para cada observação.

- **generate <nome var>=<expressão>**

Ex:

- **generate id=_n**

Gera uma nova variável *id* na qual cada indivíduo terá um número de identificação que será o mesmo que a observação índice.

Ou

Data → Create or change variable → Create new variable

- **gen porcentopeso=((pesof-pesoin)/pesoin)*100**

Gera uma nova variável *porcentopeso* que assumirá valor faltante se **pesoin** ou **pesof** for valor faltante ou será igual à porcentagem de aumento ou de diminuição de peso em relação ao peso inicial.

- **gen aumentosistolica=0 if sistolica<0**

Cria a variável *aumentosistolica* que categorizará os indivíduos entre os que tiveram aumento ou diminuição da pressão sistólica durante o período de observação. O valor 0 indicará diminuição da pressão. Os indivíduos restantes serão codificados como valores faltantes, ".".

O comando **replace** funciona como o comando **generate,** com a diferença que permite que uma variável já existente seja alterada.

- **replace aumentosistolica =1 if sistolica>=0**

  Modifica os valores faltantes para 1 se *sistolica* maior ou igual a 0.

Ou

Data → Create or change variable → Change contents of variable

Gerando variáveis indicadoras (dummy)*:*

A variável droga é categorizada em 1, 2, 3 e 4. O comando:

- **tab droga, gen(droga)**

Ou

Data→Create or change variable → Other variable creation commands →Create indicators variable

gera 4 variáveis *dummy*: droga1, droga2, droga3 e droga4 de tal forma que droga1 terá valores iguais a 1 quando a droga utilizada for a 1 e 0 se a droga utilizada for 2, 3 ou 4. A variável droga2 terá valores iguais a 1 quando a droga utilizada for a 2 e 0 se a droga utilizada for 1, 3 ou 4. E assim será para as variáveis droga3 e droga4.

Variáveis indicadoras terão aplicação, por exemplo, na construção de gráficos de pizza e análise de regressão.

Comando **egen:**

O comando **egen** pode ser função de muitas variáveis simultaneamente.

- **egen media=rowmean(pesoin-pesof) ,** onde "pesoin-pesof" significa da variável pesoin até a variável pesof.
- **egen media=rowmean(pesoin pesointer pesof)**

Ou

Data → Create or change variable → Create new variable (extend)

Cria a nova variável chamada media e calcula a média de peso para cada indivíduo usando as variáveis que existem de peso inicial até peso final. Os valores faltantes são ignorados.

rowmean trabalha nas linhas**.**

- **egen famsal=mean(salario),by(familia)**

Cria uma nova variável e calcula a média da variável **salario** para o conjunto de valores iguais de familia**.**

mean trabalha na coluna da variável.

Uma variável existente pode ser retirada do banco de dados usando o comando **drop.**

- drop salário

Ou

Data → Variables utilities → **keep or drop** variables or observations

Pode-se utilizar, também, o comando **keep <var>**, onde se deve listar as variáveis que devem permanecer no banco de dados.

SALVAR O ARQUIVO - pelo menu ou pelo comando:

- *File→Save as→ c:/cursostata/sistolica.dta*

FECHAR O ARQUIVO LOG - pelo ícone ou pelo comando:

- *log close*

2.3 - Mudando a forma de apresentação dos dados

Abrir o arquivo c:\cursostata\calorias1.dta.

Neste arquivo, para um mesmo indivíduo, são obtidas duas ou mais informações:

- **list**

	id	cal1	cal2	sexo
1.	1	2300	2500	1
2.	2	2400	3200	1
3.	3	2400	3600	1
4.	4	3200	3500	2
5.	5	3000	3200	2
6.	6	3000	3500	2
7.	7	2564	3589	1
8.	8	2600	2785	1
.	.	.	.	.
.	.	.	.	.
19.	19	3800	3500	1
20.	20	2980	2851	2

Esta forma de apresentação dos dados é denominada **wide**. A forma de apresentação dos dados pode ser mudada para o formato **long**, utilizando o comando

- **reshape long cal, i(id) j(consulta)**

Ou

Data → Create or change variable → Other variable transformation commands → Convert data between wide and long

- **list**

	id	consulta	cal	sexo
1.	1	1	2300	1
2.	1	2	2500	1
3.	2	1	2400	1
4.	2	2	3200	1
5.	3	1	2400	1
6.	3	2	3600	1
7.	4	1	3200	2
8.	4	2	3500	2
9.	5	1	3000	2
10.	5	2	3200	2
11.	6	1	3000	2
12.	6	2	3500	2
.	.	.	.	.
.	.	.	.	.
39.	20	1	2980	2
40.	20	2	2851	2

É necessário que o arquivo em formato wide tenha uma variável de identificação e que a(s) variável(is) que se repete(m) no tempo tenha(m) o mesmo prefixo no seu nome. O comando *reshape long* gera uma nova variável que identifica o número da observação. As variáveis que não se repetem no tempo como *sexo* mantêm o mesmo valor para cada observação da mesma unidade de observação (indivíduo, família, animal, etc).

Para reverter ao formato anterior (**wide**)

- **reshape wide cal, i(id) j(consulta)**
- **list**

Ou

Data → Create or change variable → Other variable transformation commands → Convert data between wide and long

reshape long <raiz(es) da(s) variável(is) periódica(s)>, i(<identificação do indivíduo>) j(<nome para a parte numérica da(s) variável(is) periódica(s)>)

Duas variáveis vistas nas duas ocasiões cada (cal1 cal2, var1 var2):

- **list,clean**

	id	cal1	cal2	sexo	var1	var2
1.	1	2300	2500	1	23	25
2.	2	2400	3200	1	24	32
3.	3	2400	3600	1	24	36
4.	4	3200	3500	2	32	35
5.	5	3000	3200	2	30	32
15.	15	3589	3600	1	35.89	36
16.	16	4001	2960	2	40.01	29.6
17.	17	2030	1990	2	20.3	19.9
18.	18	2451	2601	1	24.51	26.01
19.	19	3800	3500	1	38	35
20.	20	2980	2851	2	29.8	28.51

- **reshape long cal var, i(id) j(consulta)**
- **list,clean**

	id	consulta	cal	var	sexo
1.	1	1	2300	23	1
2.	1	2	2500	25	1
3.	2	1	2400	24	1
4.	2	2	3200	32	1
5.	3	1	2400	24	1
6.	3	2	3600	36	1
7.	4	1	3200	32	2
8.	4	2	3500	35	2
9.	5	1	3000	30	2
10.	5	2	3200	32	2

2.5- Junção de bancos de dados

O arquivo que está aberto (calorias1.dta) é denominado mestre.

Objetivo 1: Acoplar os dados de um segundo banco ao final do banco mestre, como em continuação deste. Não precisa ter necessariamente as mesmas variáveis.

- **append using <arquivo>**

Ex:

- **append using c:\cursostata\calorias2.dta**

id	cal1	cal2	sexo
1	2300	2500	1
2	2400	3200	1
3	2400	3600	1
4	3200	3500	2
5			

20

Banco Mestre

↓

id	cal1	cal2	sexo	idade
21	2560	2001	1	45
22	2330	2064	1	42
23	2648	2542	1	36
24	2900	2981	2	35
25				

40

Banco 2

Ou

Data →Combine dataset →Append datasets

Salvar como c:/cursostata/calorias12.dta

Objetivo 2: unir lado a lado dois bancos de dados que contenham informações correspondentes à mesma unidade de observação (indivíduo, família, animal, etc). É necessário que os bancos tenham uma variável de identificação (com a mesma sintaxe) e que esteja ordenado por esta variável.

- **merge <variável de identificação> using <arquivo>**

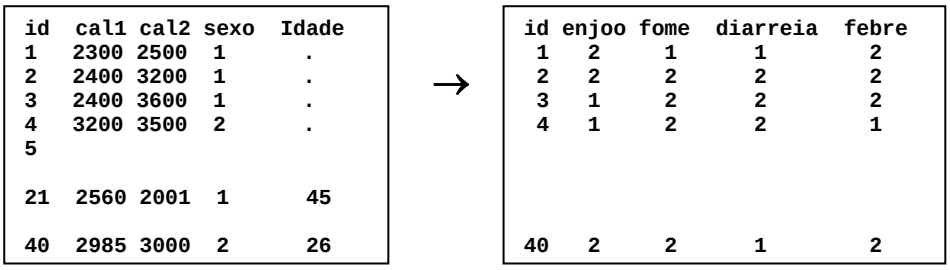
Ex:

- **sort id**
- **save, replace**

Abrir o segundo banco c:\cursostata\sintomas

- **sort id**

- **save, replace**
- **use c:\cursostata\calorias12.dta**
- **merge id using c:\cursostata\sintomas**



id	cal1	cal2	sexo	Idade
1	2300	2500	1	.
2	2400	3200	1	.
3	2400	3600	1	.
4	3200	3500	2	.
5				
21	2560	2001	1	45
40	2985	3000	2	26

→

id	enjoo	fome	diarreia	febre
1	2	1	1	2
2	2	2	2	2
3	1	2	2	2
4	1	2	2	1
40	2	2	1	2

Banco Mestre ***Banco 2***

Ou Data → Combine dataset → Merge datasets*

**** Ordenar pela variável de identificação antes de realizar este comando via menu e selecionar a variável comum aos dois bancos no menu.***

O comando *merge* gera uma variável *_merge* com os códigos:

1- dados faltantes no banco 2

2- dados faltantes no banco mestre

3- união de dados realizada com sucesso

Salvar o banco de dados com o nome: inteiro.dta

---

3. Descrição de dados

3.1- Gráficos

Alguns tipos de gráficos que o Stata 9/10 executa, e seus comandos, estão apresentados na tabela abaixo:

COMANDO	TIPO DE GRÁFICO
graph box	**box–plots**
twoway scatter, line, lfit, qfit	**diagrama de dispersão, regressão, linhas**
graph matrix	**matriz de diagrama de dispersão**
histogram	**histograma**
qnorm	**gráfico de quantis para normal**
(ladder), qladder, gladder	**gráficos de diagnósticos para normal**
graph pie	**gráfico de setores circulares (pizza)**
graph bar	**gráfico de barras**

Abrir o banco de dados c:/cursostata/ sistolicainic.dta

A sintaxe básica para a elaboração de gráficos é:

- **graph <tipo> <var>, options**

Há várias opções para gráficos, muitas vezes há que usar o *Help* ou o manual. Para confeccionar gráficos, o menu ajuda muito, pois já traz as opções, que não precisam ser digitadas na linha de comando.

Os gráficos não podem ser copiados no arquivo log. Deve-se abrir um arquivo .doc previamente; obtido o gráfico, ***no menu***, clicar em **Edit → Copy graph** e depois **colar** no doc. Os gráficos também podem ser salvos com extensão .gph (**File**

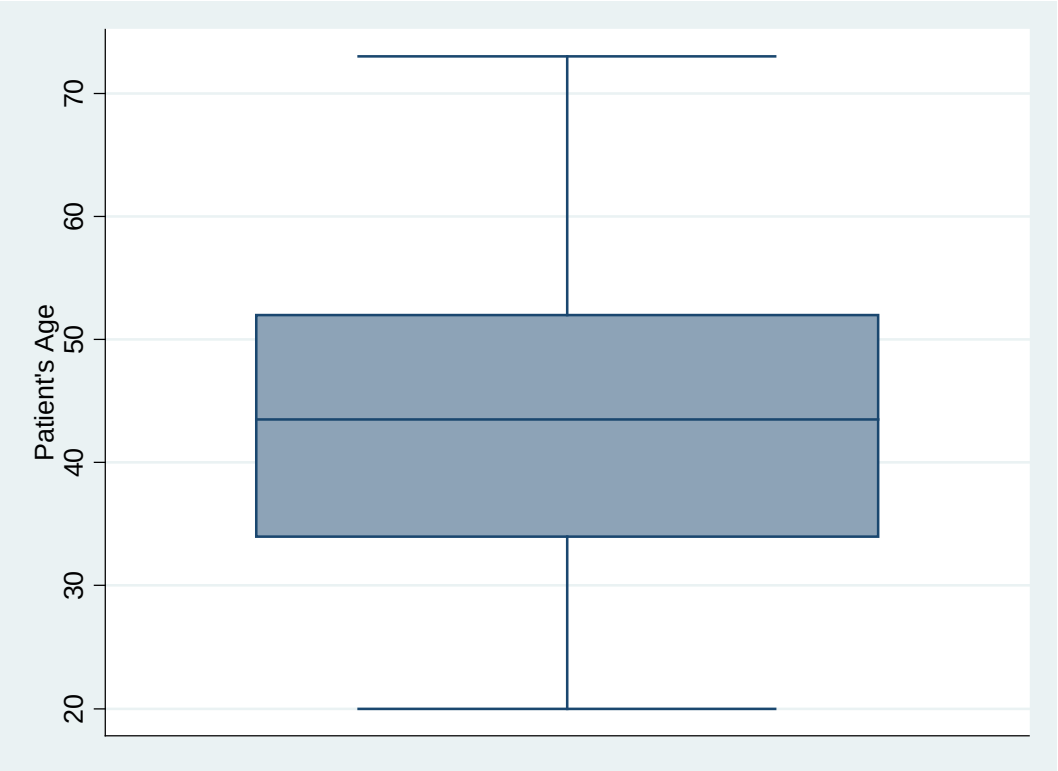
Save Graph), porém só poderão ser abertos novamente no Stata e não em editor de texto ou em arquivo .doc.

Boxplot

- **graph box idade**

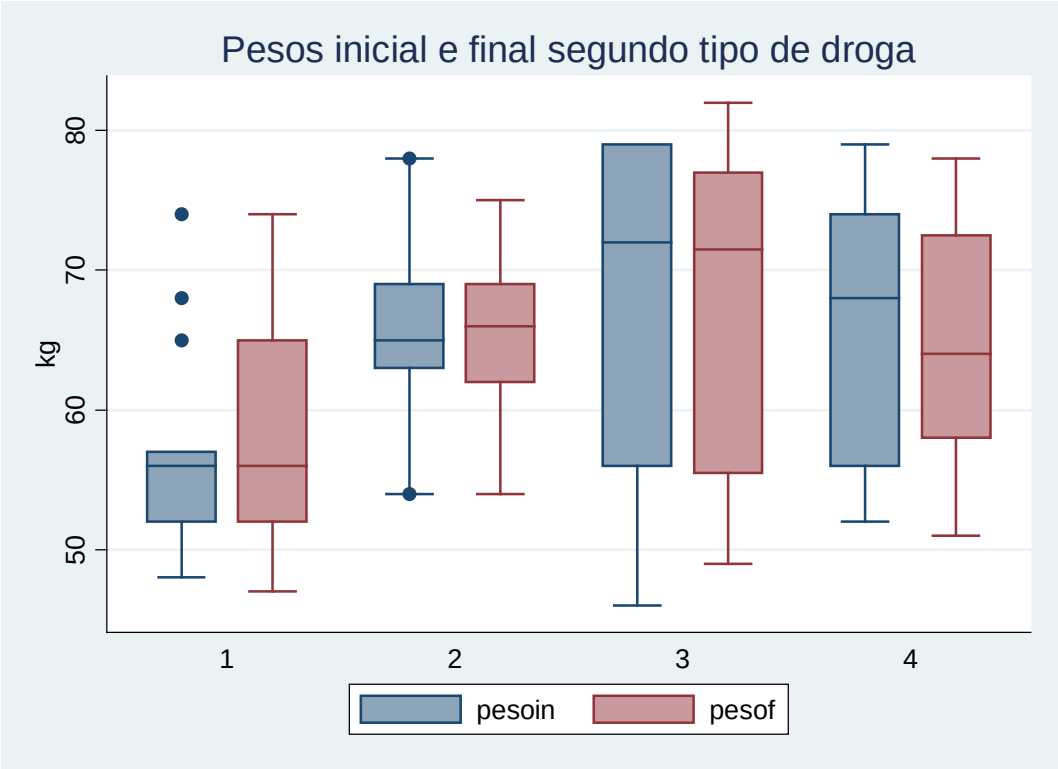
Ou Graphics ⟶ Box plots

Produz um boxplot da variável idade



- **graph box pesoin pesof, over(droga) ytitle(kg) title(Peso inicial e final segundo tipo de droga)**

Cria um boxplot, contendo pesoin e pesof, segundo tipo de droga em um mesmo conjunto de eixos ortogonais. A opção *ytitle* define o nome do eixo y e *title* o nome do gráfico.



- **graph hbox pesoin pesof, by(droga)**

Fornece um boxplot na horizontal (*hbox)* para cada categoria de droga, em dois conjuntos de eixos ortogonais independentes.

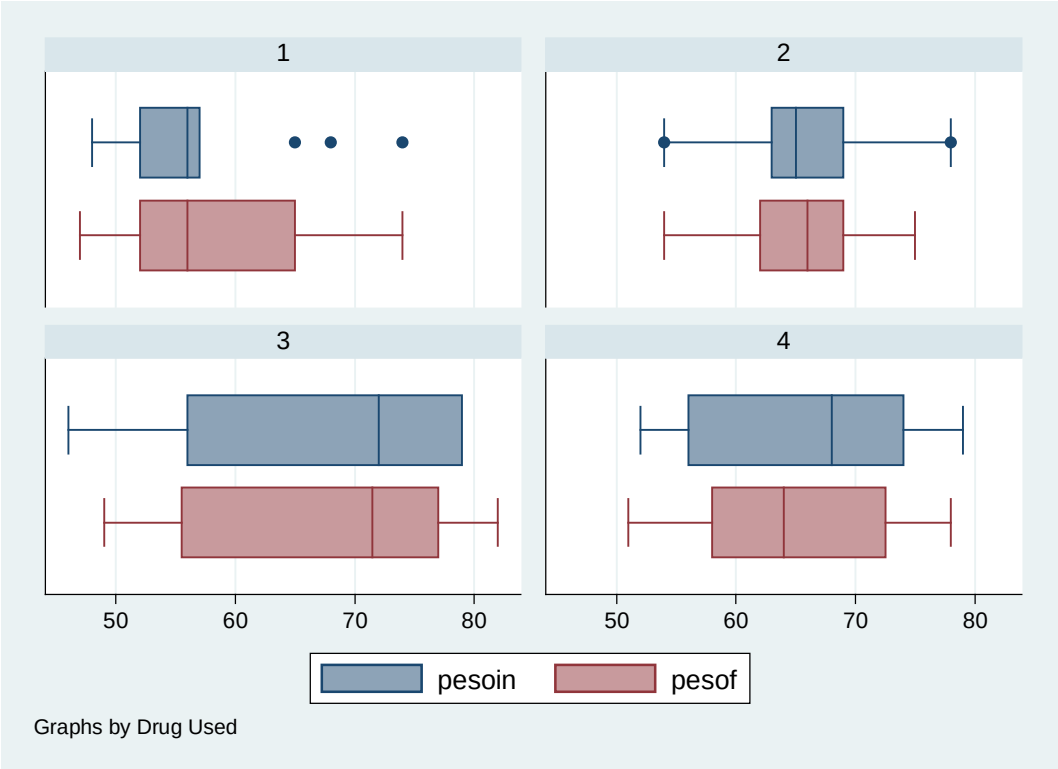


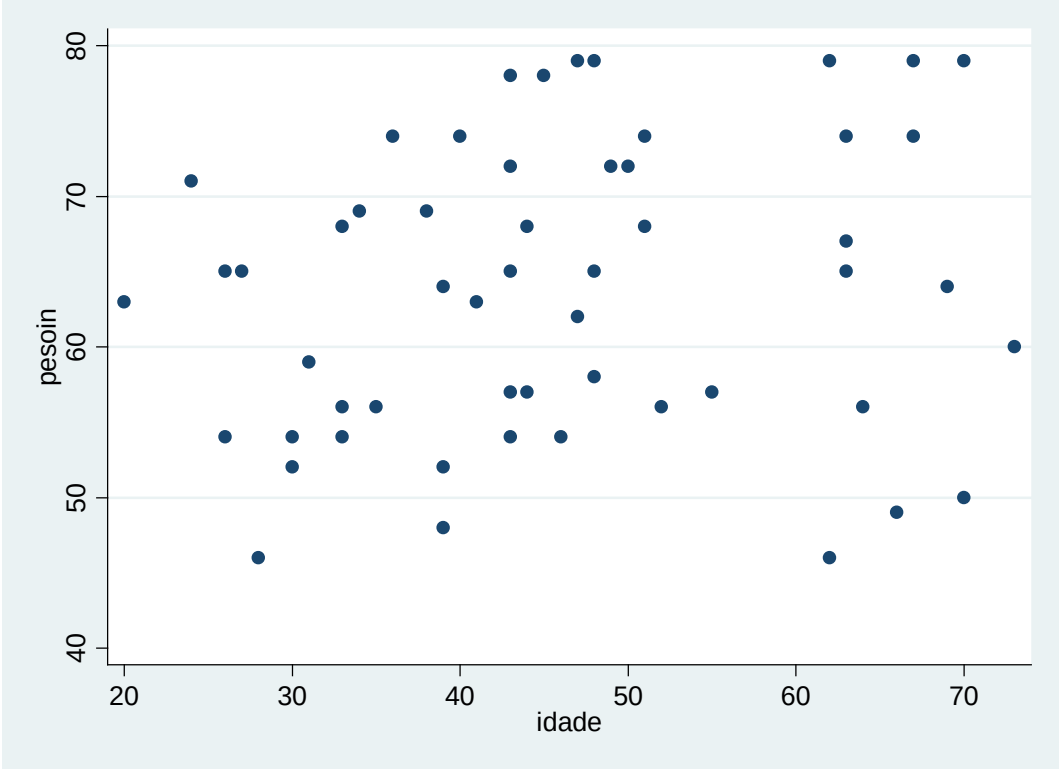
Diagrama de dispersão

- **scatter pesoin idade,xtitle(idade)**

Fornece um diagrama de dispersão de idade e pesoin

Ou

Graphics ⟶ twowaygraph (scatter plot, line, etc.)



- **twoway (lfit pesoin idade,lwidth(thick)lcolor(blue)) (scatter pesoin idade, msymbol(square) msize(vlarge) mlabel(droga)), title (Peso inicial X Idade) xtitle(idade) ytitle(peso inicial) legend(col(2)lab(1 "valores ajustados")lab(2 "pesoinicial"))**

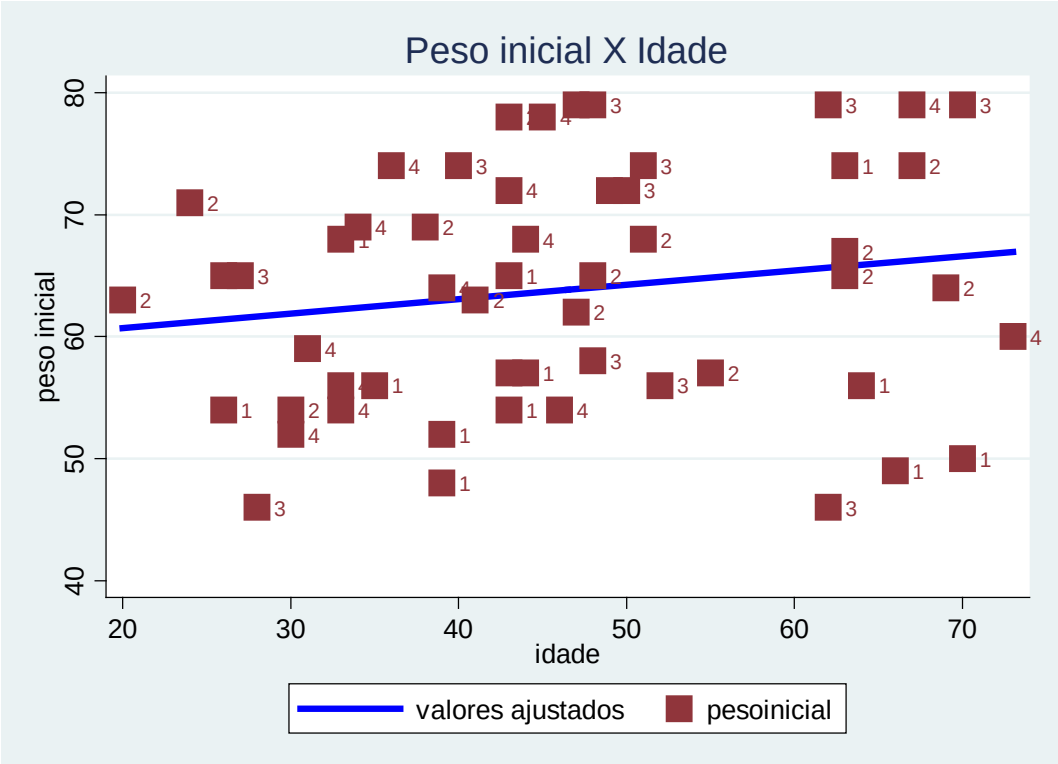
- **twoway lfit pesoin idade,lwidth(thick)lcolor(blue)||scatter pesoin idade, msymbol(square) msize(vlarge) mlabel(droga)|| , title (Peso inicial X Idade) xtitle(idade) ytitle(peso inicial) legend(col(2)lab(1 "valores ajustados")lab(2 "pesoinicial"))**

Ou

Graphics ⟶ Overlaid twoway graphics

lfit é um gráfico que mostra a reta de regressão. O comando twoway pode construir dois gráficos sobrepostos. Os subcomandos são separados por parênteses ou por barra dupla || e cada um pode apresentar opções especificas. O subcomando *lfitci* coloca o intervalo de 95% de confiança para a linha da regressão.

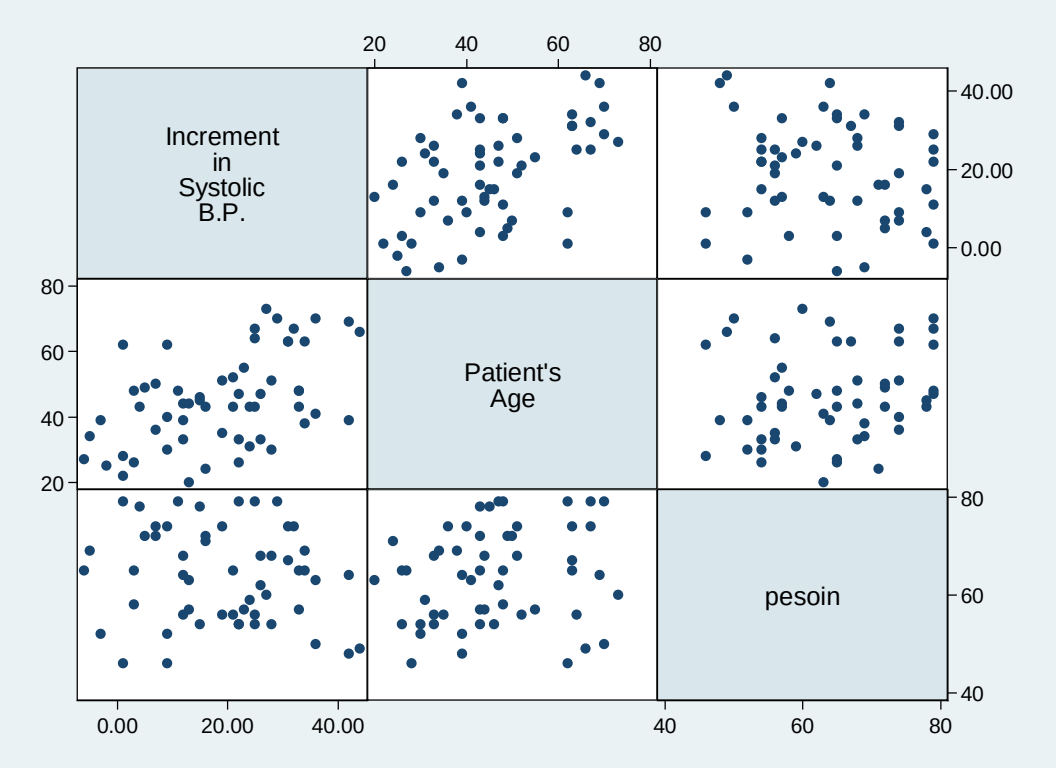
- **scatter pesoin idade,xtitle(idade)||lfit pesoin idade, lcolor(blue)**



Matriz de diagramas de dispersão

Constrói figura com gráficos de dispersão dos pares de variáveis.

- **graph matrix sistolica idade pesoin**



Histograma

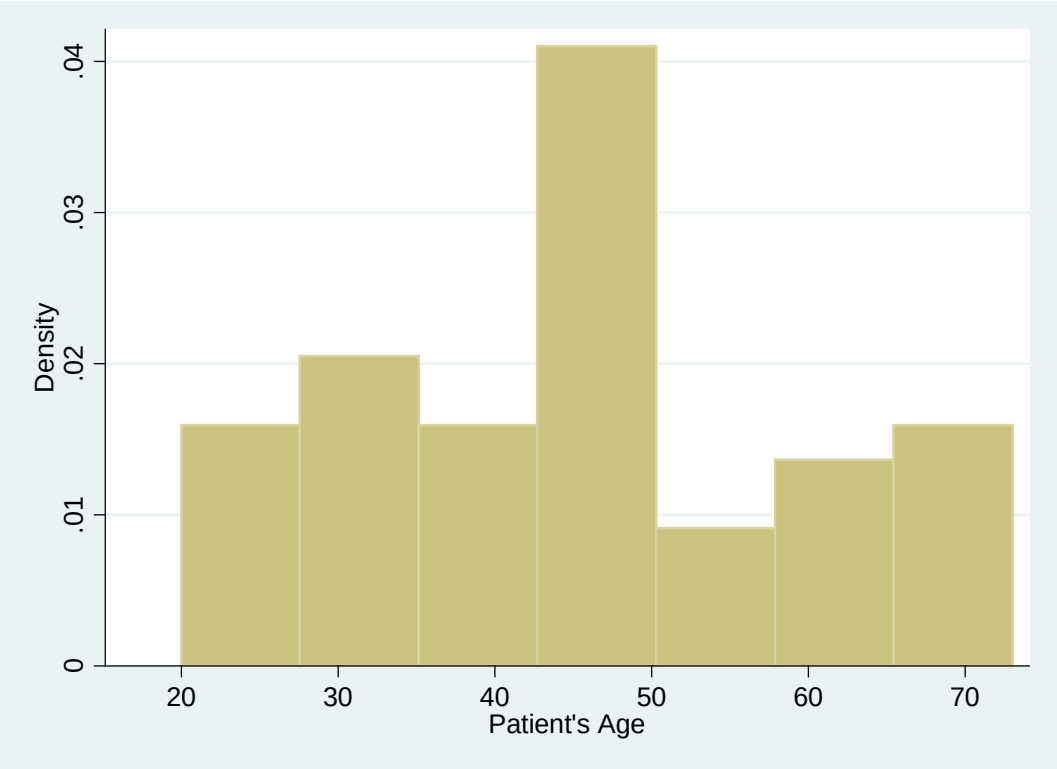
O histograma não precisa do comando *graph* antes:

- **histogram idade**

Ou

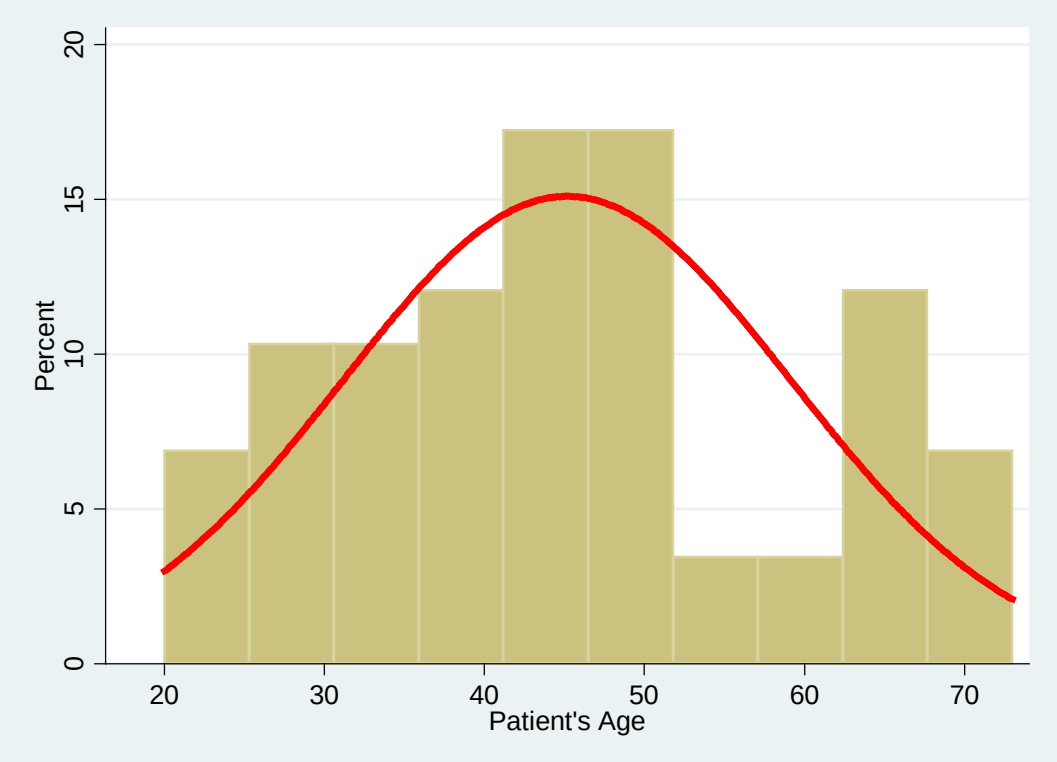
Graphics ⟶ Histogram

Desenha um histograma da variável idade.


Density
0
.01
.02
.03
.04
20
30
40
50
60
70
Patient's Age

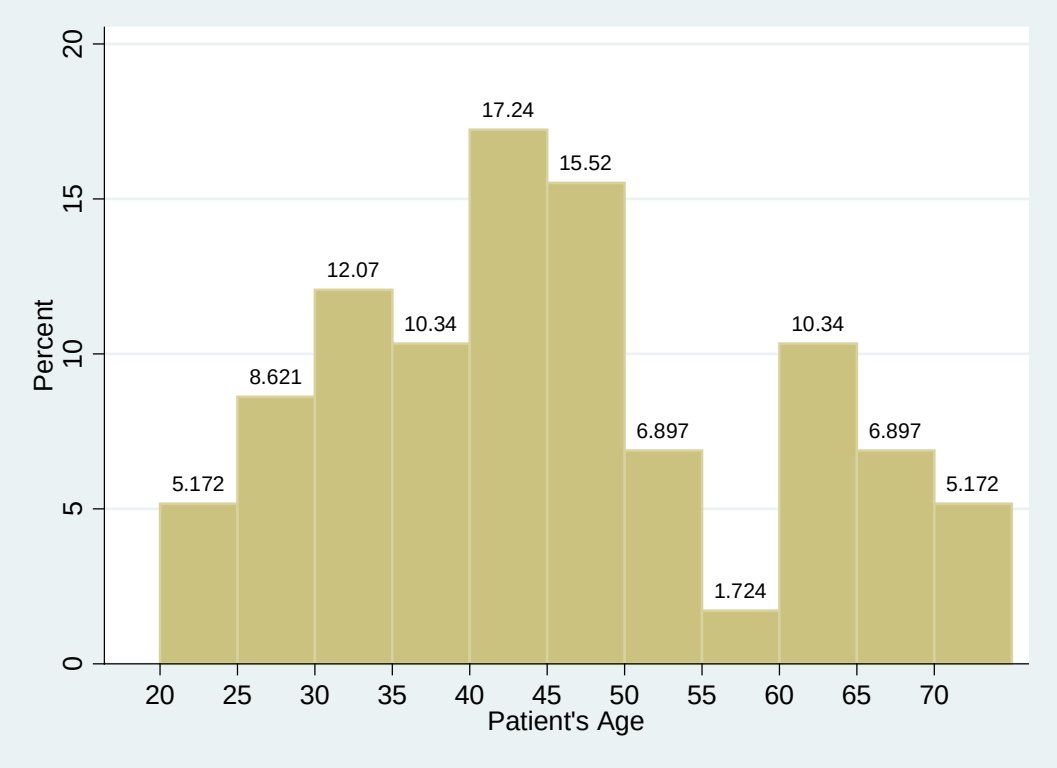
- **histogram idade,bin(10) percent norm normopts(lcolor(red) lwidth(thick))**

Desenha um histograma da variável **idade** em 10 intervalos de classe. O número de intervalos pode variar, de acordo com os dados. *percent* está definindo que o eixo y deve ser representado pelas porcentagens de unidades de observação. *norm* superpõe o desenho da curva normal com a média e o desvio padrão dos respectivos dados.



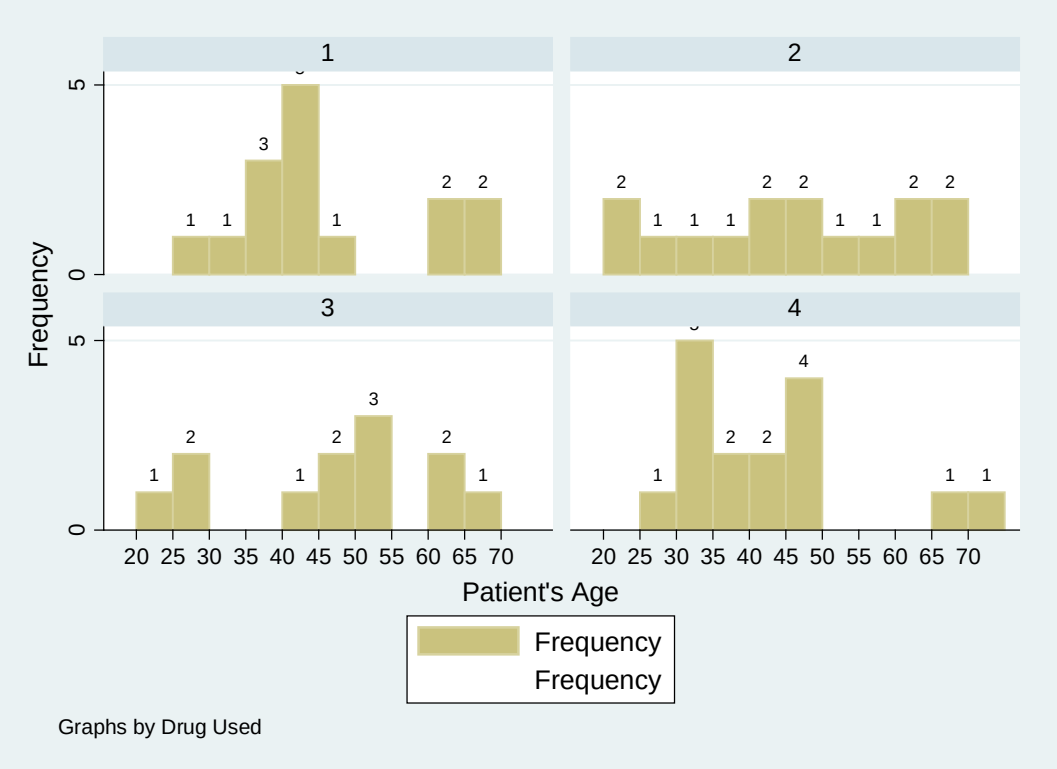
- **histogram idade, width(5) addlabels xlabel(20(5)70) percent**

A opção *width* estabelece o tamanho do intervalo de cada coluna (não pode ser usado junto com a opção *bin*). A opção *addlabels* coloca a legenda do número de cada barra e *xlabel* define os rótulos do eixo x (mínimo = 20, com intervalo de 5 anos e máximo = 70).



- **histogram idade, width(5) addlabels xlabel(20(5)70) by(droga) frequency**

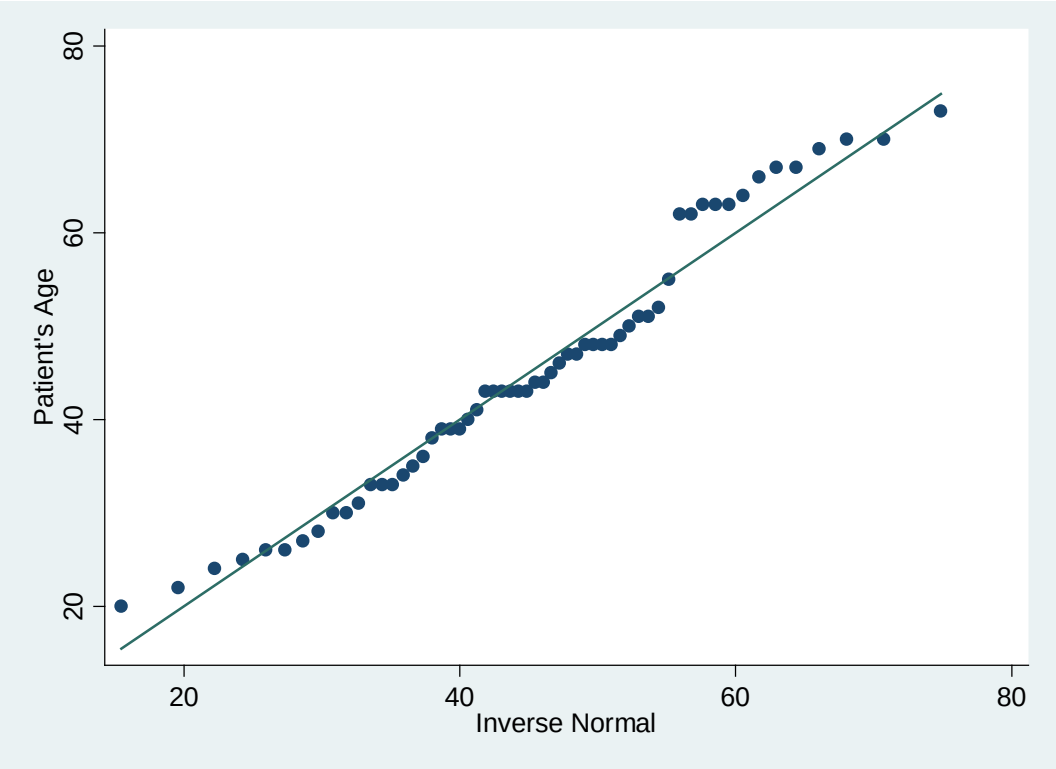
A opção *by(droga)* faz com que seja desenhado um histograma para cada tipo de droga.



Quantis de variável contínua vs quantis de distribuição normal (diagnostic plots)

Para orientar sobre "normalidade" de uma variável pode-se u**qnorm** ou o conjunto ladder, gladder, qladder**:**

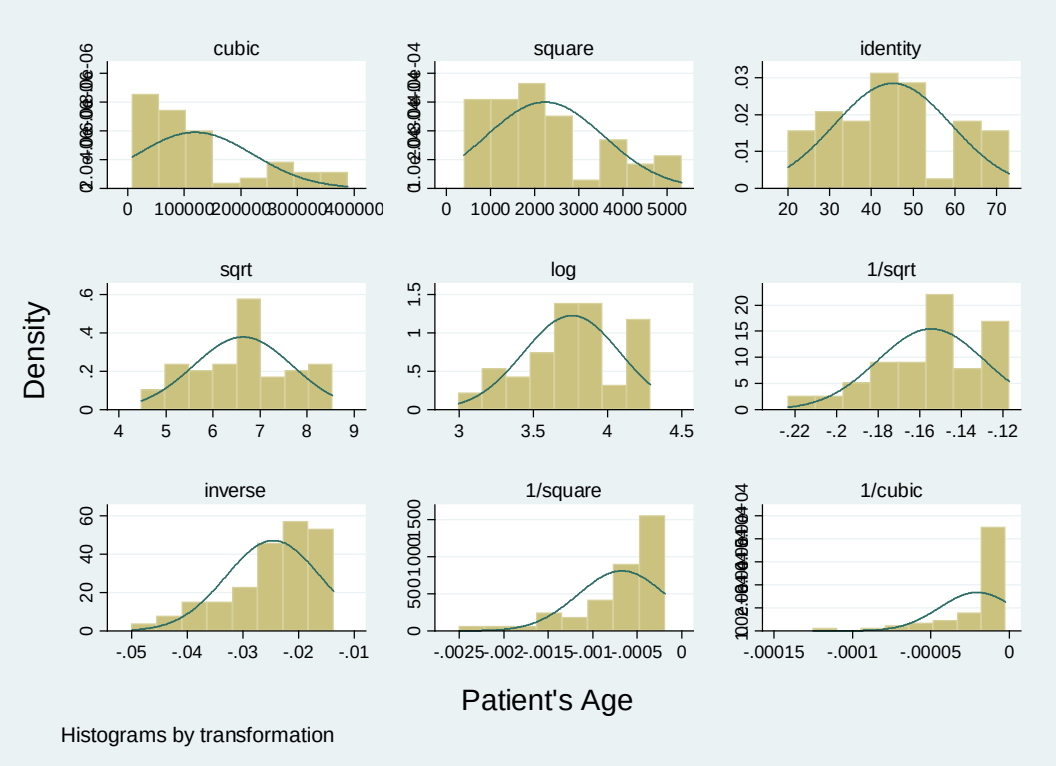
- **qnorm idade**



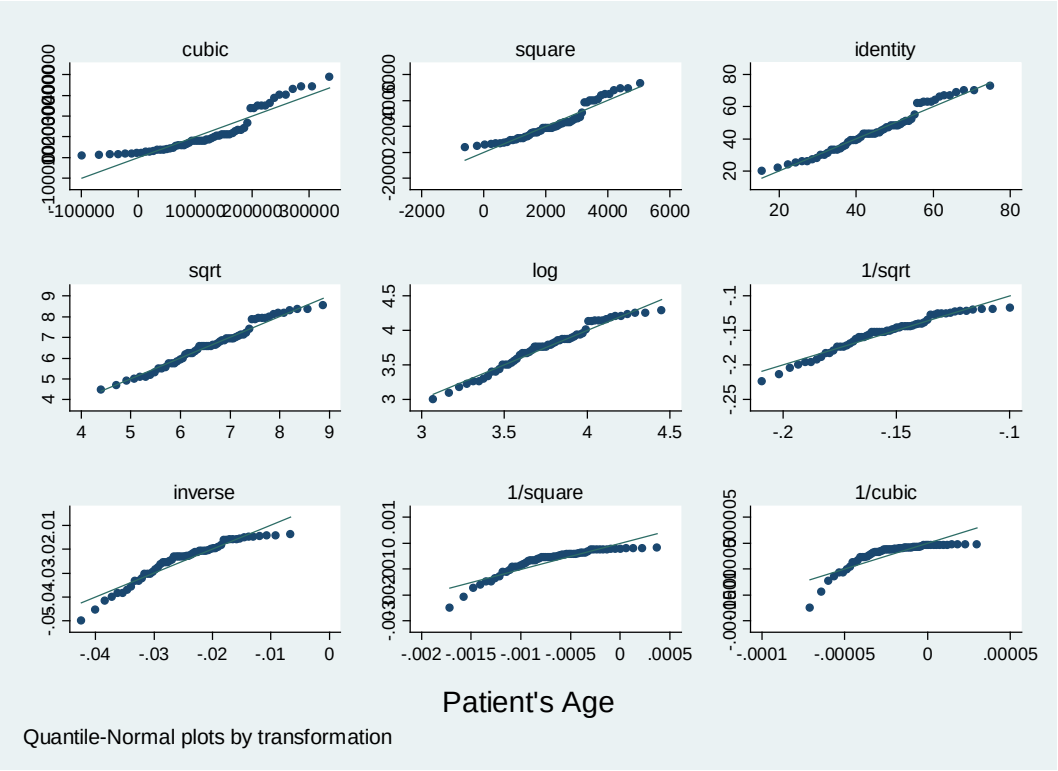
- **ladder idade**

Transformation	formula	chi2(2)	P(chi2)
cubic	idade^3	8.91	0.012
square	idade^2	5.62	0.060
identity	idade	3.87	0.144
square root	sqrt(idade)	2.72	0.256
log	log(idade)	2.19	0.335
1/(square root)	1/sqrt(idade)	4.42	0.110
inverse	1/idade	8.69	0.013
1/square	1/(idade^2)	20.55	0.000
1/cubic	1/(idade^3)	32.64	0.000

- **gladder idade**



- **qladder idade**



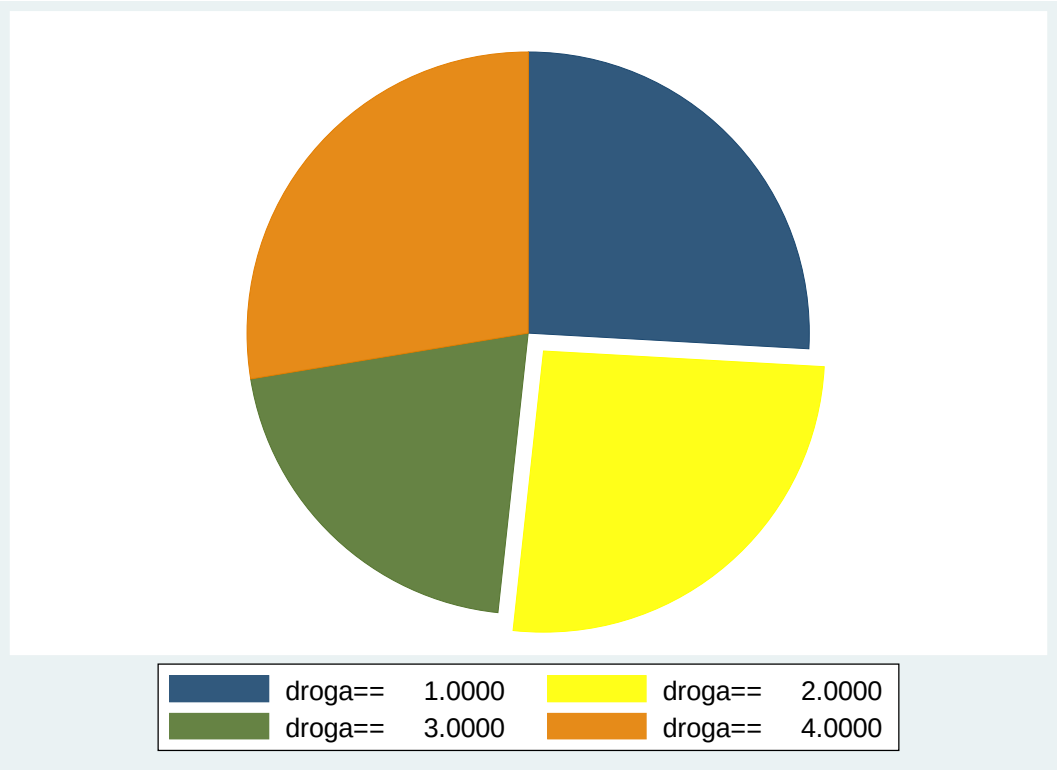
Pizza

Desenhar um gráfico de pizza tendo criado uma variável *dummy:*

- **tab droga, gen(droga)**
- **graph pie droga1 droga2 droga3 droga4 , pie(2,explode color(yellow))**

Desenhar usando a opção "over":

- **graph pie, over(droga) pie(2,explode color(yellow))**

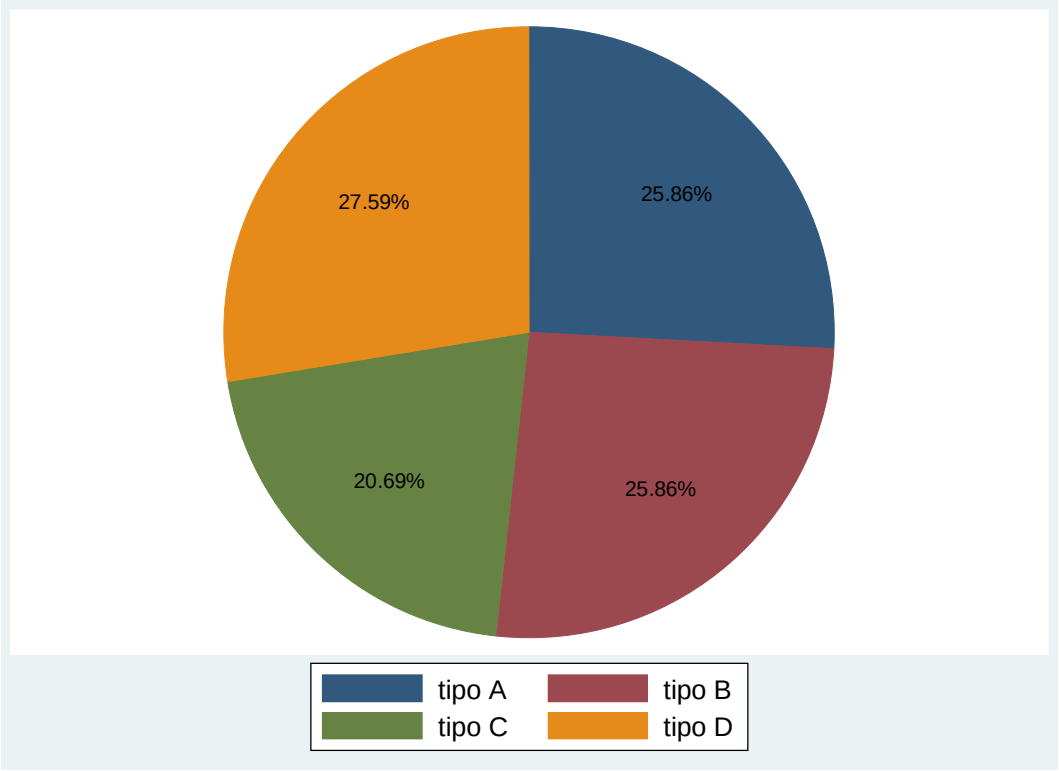


Ou

Graphics ⟶ Pie chart

- **graph pie droga1 droga2 droga3 droga4, plabel(_all percent) legend(order(1 “Tipo A” 2 “Tipo B” 3 “Tipo C” 4 “Tipo D”))**

Plabel faz com que no gráfico apareçam os rótulos de porcentagem de todos os pedaços. *Legend* formata a legenda.

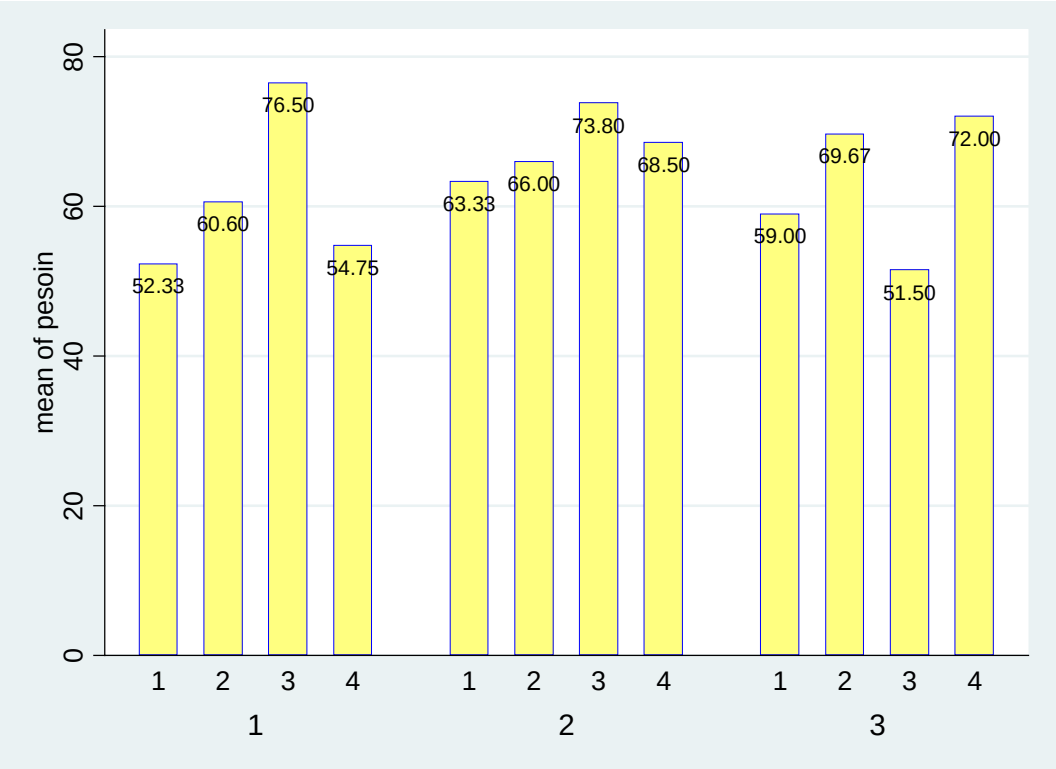


Para mudar as características dos gráficos, com o *mouse* selecionar *Graphics* na barra de menu e em seguida *Graph Preferences* ou *Change scheme/size*

Gráfico de barras

Há um grande número de possibilidades.

- **graph bar (mean) pesoin, over(droga) bla-bel(bar,position(inside)format(%5.2f)) intensity(50) over(doenca) bar(1,lcolor(blue) fcolor(yellow))**



3.2 – Tabelas e resumo dos dados

Criar um arquivo .log c:\cursostata\feminic.log

Abrir o banco de dados c:\cursostata\feminic.dta

Os dados constituem uma amostra de 118 pacientes psiquiátricos, do sexo feminino e estão disponíveis em D.J. Hand et al. ***A Handbook of Small Data Sets***. Chapman &Hall, London, 1994. As variáveis estudadas são:

- **age**: idade em anos
- **iq**: escore de inteligência
- **anxiety**: ansiedade (1=nenhuma, 2=leve, 3=moderada, 4=severa)
- **depress**: depressão (1=nenhuma, 2=leve, 3=moderada)
- **sleep**: dorme normalmente? (1=não, 2=sim)
- **sex**: perdeu interesse em sexo? (1=não, 2=sim)
- **life**: tem pensado em acabar com sua vida? (1=não, 2=sim)
- **weight**: mudança de peso durante os últimos 6 meses (em libras)
- **id**: número de identificação

Preparação do banco de dados antes de começar a análise:

Objetivo	**Comandos**
Verificar quais são as variáveis que compõem o banco de dados	describe ou desc
Substituir o valor -99 pelo símbolo **.**, para indicar ausência de informação	mvdecode _all, mv(-99)
Recodificar variáveis	recode sleep 1=0 2=1 {0=não;1=sim} recode sex 1=0 2=1 {0=não;1=sim} recode life 1=0 2=1 {0=não;1=sim}
Criar rótulos (*labels*) para as variáveis	label define sn 0 não 1 sim label values sex sn label val sleep sn label val life sn
Criar um rótulo para a variável **weight**	label variable weight "mudou peso nos últimos 6 meses"
Criar rótulo para a variável **life**	label variable life "pensou em terminar com sua vida recentemente?"
Criar uma variável **ageg** contendo a variável age em intervalos de classes de 5 anos	gen ageg=age recode ageg 25/29=1 30/34=2 35/39=3 40/44=4 45/49=5 label define id 1 "25-29" 2 "30-34" 3 "35-39" 4 "40-44" 5 "45-49" label val ageg id

Tabela de frequência

- **tabulate age**

Ou Statistics ⟶ Summaries, tables & tests ⟶ Tables ⟶ One way tables

Não é possível solicitar mais de uma tabela na mesma linha de comando, por exemplo, *tab age iq*. Quando esse for o objetivo é necessário utilizar o comando:

- **tab1 ageg iq**

Ou Statistics ⟶ Summaries, tables & tests ⟶ Tables ⟶ Multiple one way tables

Tabela de contingência

- **tab life sex**

Ou

Statistics ⟶ Summaries, tables & tests ⟶ Tables ⟶ Two way tables with measures of association

- **sort sleep**
- **by sleep: tab life sex, chi2 col row**

A opção *chi2* mostra o resultado do teste qui-quadrado. As opções *row* e *col* mostram as freqüências relativas na linha e coluna.

Tabela contendo resumo de outras variáveis

- **table life, contents(mean iq sd iq)**

Ou

Statistics ⟶ Summaries, tables & tests ⟶ Tables ⟶ Table of summary statistics (table)

Resumo de variáveis

- **sum age, detail**

A opção *detail* mostra também os percentis

Ou

Statistics ⟶ Summaries, tables & tests ⟶ Summary statistics ⟶ Summary statistics ⟶ **display additional statistics.**

Salvar o banco de dados com o nome "fem.dta"

4. Análise de dados epidemiológicos

Banco de dados: feminic.dta

Comparação de médias:

Para comparar as variáveis quantitativas entre grupos pode-se utilizar o teste *t de "Student"* que assume que as observações nos dois grupos são independentes; as amostras foram retiradas de populações com distribuição normal, com mesma variância. Um teste alternativo, não paramétrico, que não necessita destas pressuposições, é o teste *U* de *Man-Whitney*. Para mais de dois grupos independentes, utiliza-se a análise de variância (ANOVA) oneway; a análise correspondente na estatística não paramétrica é o teste de *Kruskal-Wallis.*

Coeficiente de correlação:

É possível calcular correlações entre variáveis contínuas. Se se quiser testar se o coeficiente de correlação de *Pearson* é estatisticamente diferente de zero, o Stata apresenta um teste que pressupõe que as variáveis são normais bivariadas. Se esta pressuposição não for satisfeita, pode-se utilizar a correlação de postos de *Spearman*. Se as variáveis forem categóricas é possível utilizar a estatística de *Kendall* como medida de associação.

Associação entre variáveis:

Para as variáveis qualitativas nominais pode-se utilizar o teste qui-quadrado, de *Pearson*.

O Stata realiza alguns testes estatísticos sem a necessidade de haver um banco de dados. São comandos que terminam com ***i*** *ou pelo menu* ***Statistics*** *que tenham a palavra* ***calculator.***

4.1 – Teste de hipóteses para uma, duas ou mais médias e intervalos de confiança

Intervalo de 95% de confiança de média

- **ci weight**

Ou

Statistics → Summaries, tables & tests → Summary statistics → Confidence Intervals

Intervalo de 95% de confiança para uma dada amostra, média e desvio padrão

- **cii 100 2 2.5**

Amostra=100; Média observada=2; Desvio padrão populacional=2,5

Ou

Statistics → Summaries, tables & tests → Summary statistics → Normal CI calculator

Teste de duas médias (t de "Student") entre grupos

- **ttest weight, by (life)**
- **ttest weight1=weight 2 ;(teste pareado)**
- **ttest diferençaweight1menosweight 2=0 ;(equivale ao anterior)**
- **ttest weight=0**
- **ttest weight1=weight2,unp ;(teste não pareado)**

Ou

Statistics → Summaries, tables & tests → Classical tests of hypotheses → Group mean comparison test

Testar a hipótese de que a média observada é igual a um valor

- **ttest weight=2**

Testa se a média da variável **weight** (1,58) é igual à média populacional 2

Ou

Statistics → Summaries, tables & tests ⟶ Classical tests of hypotheses ⟶ One sample mean comparison test

Teste de duas médias pelo método não paramétrico (Mann-Whitney)

- **ranksum weight, by (life)**

Ou

Statistics → Summaries, tables & tests ⟶ Nonparametric tests of hypotheses⟶ Mann-Whitney two-sample ranksum test

Análise de variância com um fator (ANOVA)

- **oneway weight depress, bonferroni tabulate**

bonferroni: testes que identificam diferenças significantes;
tabulate: mostra um quadro resumo contendo a média e o desvio padrão das categorias.

- **anova weight depress**
- **loneway weight depress**

Ou

Statistics → linear models and related⟶ ANOVA ⟶ one way ANOVA ⟶ One way analyses of variance

Teste de mais de duas médias pelo método não paramétrico (Kruskal-Wallis)

- **kwallis weight, by (depress)**

Ou

Statistics → Summaries, tables & tests ⟶ Nonparametric tests of hypotheses⟶ Kruskal-Wallis rank test

4.2 – Teste de hipóteses e intervalo de confiança para proporção

Testar a hipótese de que a proporção observada é igual a um valor

Para este teste é necessário que a variável esteja codificada em 0 e 1, portanto:

- **recode life 1=0 2=1 {0=não; 1=sim}**
- **bitest life=0.5**

Testa se a proporção de pessoas que pensaram em se matar (life=1) é equivalente a 0,5 (50%).

Ou

Statistics → Summaries, tables & tests ⟶ Classical tests of hypotheses ⟶ Binomial probability test

Associação de variáveis categóricas

Teste qui-quadrado

- **tab life depress, col row chi2**

Teste exato de Fisher

- **tab life sleep, col row exact**

Ou

Statistics ⟶ Summaries, tables & tests ⟶ Tables ⟶Two way tables with measures of association

4.3 – Teste de hipóteses para correlação

Calcular a correlação de Pearson

- **corr weight iq age**

Ou

Statistics → Summaries, tables & tests ⟶ Summaries statistics ⟶ Correlations & Covariances

Calcular a correlação pelo método não paramétrico (Teste de Spearman)

- **spearman weight age**

Ou

Statistics → Summaries, tables & tests ⟶ Nonparametric tests of hypotheses⟶ Spearman´s rank correlation

5. Análise de medidas de efeito

Todos os comandos de estimação seguem a mesma estrutura em sua sintaxe:

[xi:] command depvar [model] [weights],options

A variável resposta é especificada por **depvar** e as variáveis explanatórias pelo **model**.

5.1- Regressão linear (regress)

Abrir o arquivo c:\cursostata\fem.dta

- **regress weight age**

Ou

Statistics → Linear regression and related ⟶ Linear regression

```
      Source |       SS       df       MS              Number of obs =     107
-------------+------------------------------           F(  1,   105) =   21.93
       Model |  135.142248     1  135.142248           Prob > F      =  0.0000
    Residual |   647.13383   105  6.16317933           R-squared     =  0.1728
-------------+------------------------------           Adj R-squared =  0.1649
       Total |  782.276078   106    7.379963           Root MSE      =  2.4826
------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |    .233029   .0497642     4.68   0.000     .1343559    .3317022
       _cons |  -7.158987   1.882679    -3.80   0.000    -10.89199   -3.425981
------------------------------------------------------------------------------
```

Ajusta um modelo de regressão linear de weight (variável dependente) em função de age (variável independente quantitativa contínua).

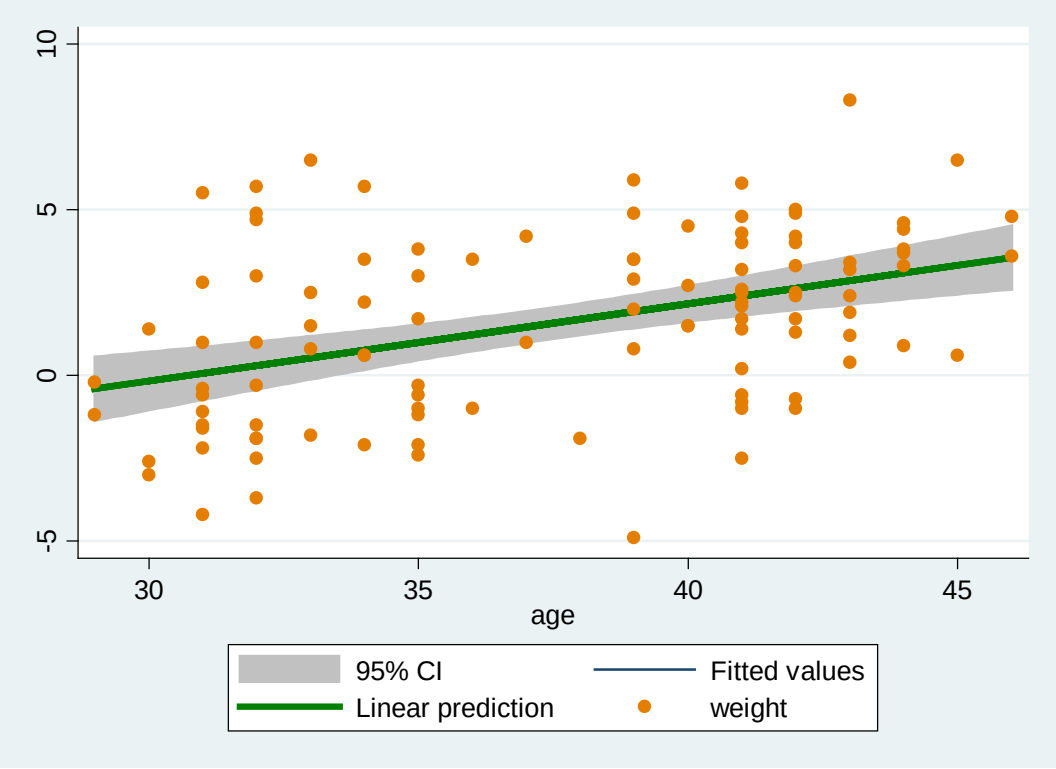
- **predict dp,stdp**
- **predict xb,xb**
- **gen low=xb-1.96*dp**
- **gen high=xb+1.96*dp**

- **twoway line xb age||scatter weight age||line low age,sort||line high age, sort**

- **twoway line xb age||scatter weight age|| line low age,sort|| line high age, sort||lfitci weight age**

- **twoway lfitci weight age||line xb age||scatter weight age|| line low age,sort||line high age, sort||**

- **twoway lfitci weight age||line xb age||scatter weight age|| rarea low high age,sort**


10
5
0
-5
30
35
40
45
age
95% CI
Fitted values
Linear prediction
weight

- **tab depress, gen(depress)**

Cria variável ***indicadora*** (*dummy*)

- **regress weight age depress2 depress3**

```
      Source |       SS       df       MS              Number of obs =     102
-------------+------------------------------           F(  3,    98) =    7.21
       Model |  137.662465     3  45.8874883           Prob > F      =  0.0002
    Residual |  623.702733    98   6.3643136           R-squared     =  0.1808
-------------+------------------------------           Adj R-squared =  0.1557
       Total |  761.365198   101  7.53826928           Root MSE      =  2.5228

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .2366617   .0533499     4.44   0.000     .1307907    .3425328
    depress2 |  -.2016017   .6107479    -0.33   0.742    -1.413611    1.010408
    depress3 |     .48972   .8145841     0.60   0.549    -1.126796    2.106236
       _cons |  -7.322033   2.108916    -3.47   0.001    -11.50711   -3.136959
------------------------------------------------------------------------------
```

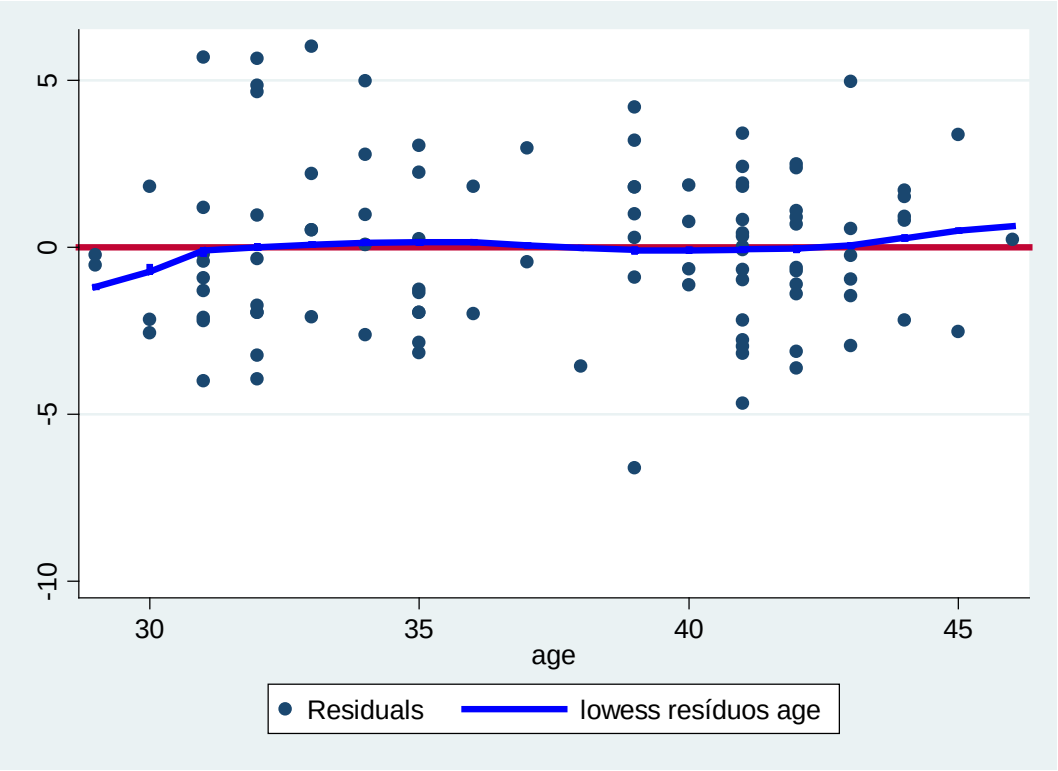
Ajusta um modelo de regressão de weight com depress2 e depress3**,** tendo depress1 como referência (variáveis dummy; a categoria referência (depress1) não é colocada no comando).

Pode-se completar a análise construindo-se gráficos, entre eles o de resíduos e o das retas de regressão ajustadas.

- **predict xb,xb**

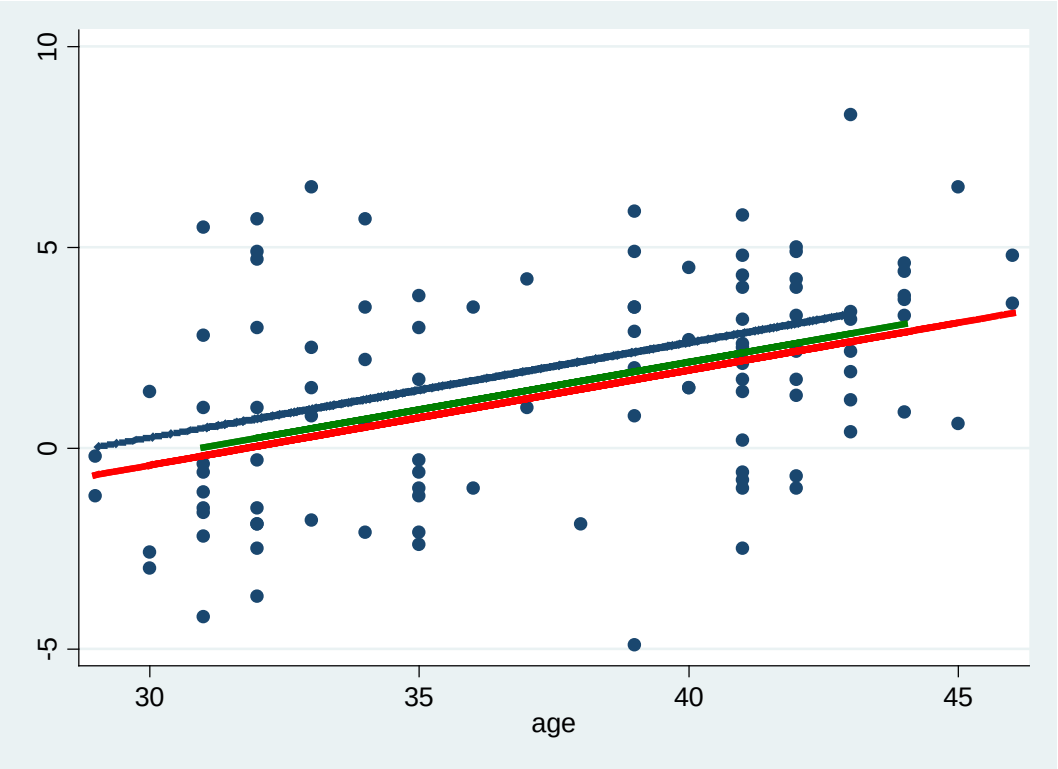
- **predict resíduos, residuals**

- **scatter resíduos age, yline(0, lwidth(thick))||lowess resíduos age, lcolor(blue) lwidth(thick)**



lowess é comando que ajusta regressão não paramétrica aos dados.

- **scatter weight age||line xb age if depress==1, lcolor(green) lwidth(thick) lpattern(solid)||line xb age if depress==2, lcolor(red) lwidth(thick) lpattern(dash)||line xb age if depress==3, lcolor(navy) lwidth(thick) lpattern(dot)||,legend(off)**



- **gen ageXdepress1=age* depress1**

(8 missing values generated)

- **gen ageXdepress2=age* depress2**

(8 missing values generated)

- **gen ageXdepress3=age* depress3**

(8 missing values generated)

São criadas três variáveis de interação.

- **reg weight age depress2 depress3 ageXdepress2 ageXdepress3**

```
------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .2165171   .1318667     1.64   0.104    -.0452363    .4782705
    depress2 |  -.3255124   5.637893    -0.06   0.954    -11.51664    10.86562
    depress3 |  -3.871144   7.219156    -0.54   0.593    -18.20106    10.45877
ageXdepress2 |   .0026154   .1472195     0.02   0.986    -.2896131    .2948438
ageXdepress3 |   .1148945   .1882496     0.61   0.543    -.2587779     .488567
       _cons |  -6.549821   5.081434    -1.29   0.201    -16.63639    3.536747
------------------------------------------------------------------------------
```

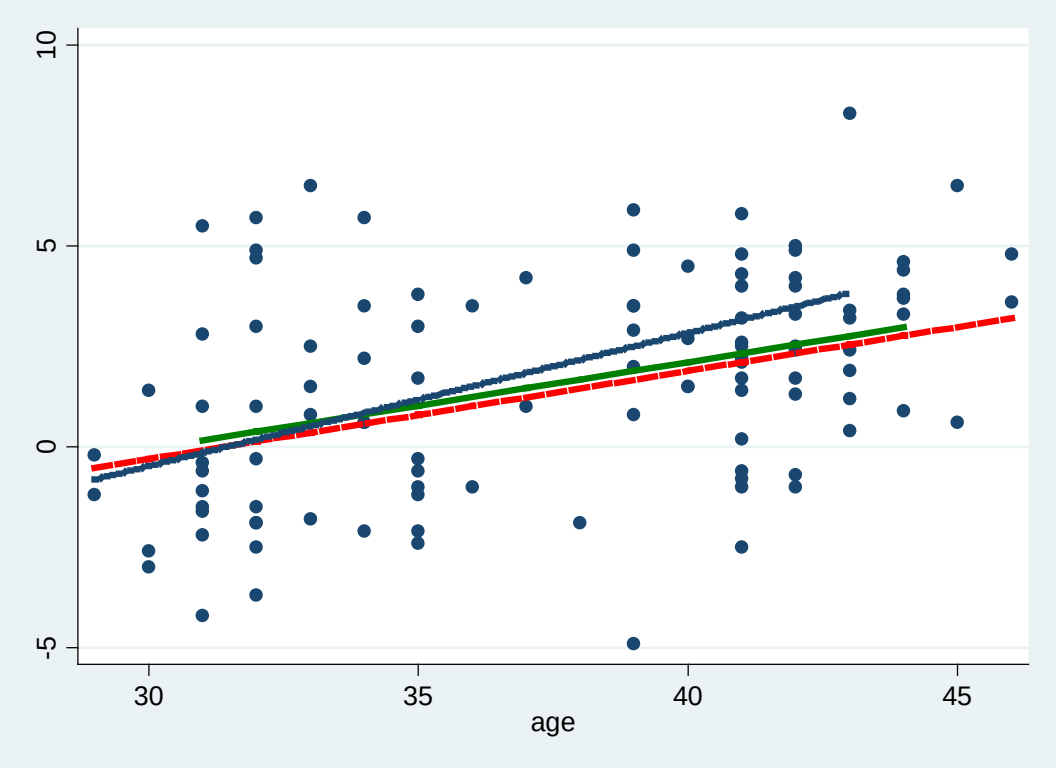
- **predict xbinter,xb**

(8 missing values generated)

- **predict residuosinter,residuals**

(16 missing values generated)

- **scatter weight age||line xbinter age if depress==1, lcolor(green) lwidth(thick) lpattern(solid)||line xbinter age if depress==2, lcolor(red) lwidth(thick) lpattern(dash)||line xbinter age if depress==3, lcolor(navy) lwidth(thick) lpattern(dot)||,legend(off)**


10
5
0
-5
30
35
40
45
age

5.2 - Regressão logística (incidência acumulada, prevalência)

Será utilizado o banco de dados originário de um ensaio clínico onde pacientes com câncer de pulmão foram alocados aleatoriamente para receber dois tipos diferentes de quimioterapia (terapia seqüencial e alternada). A variável resposta foi classificada em 4 categorias: a doença piorou (=1), não houve mudança (=2), alguma melhora (=3) e melhora total (=4). Os dados foram publicados por Holtbrugge e Schumacher (1991). A análise principal será comparar as duas terapias quanto ao desenlace.

Abrir o arquivo c:\cursostata\tumor.dta

Abrir a planilha de dados pelo *edit* ou *browse*

terapia sequencial= 0; terapia alternada= 1 (talvez a menos eficaz)

Transformando a variável resposta em uma variável dicotômica:

- **tab resposta, nol**
- **gen resultado=resposta**
- **recode resultado 1/2=1 3/4=0**

Portanto 1= mau resultado e 0= bom resultado

- **recode sexo 1=0 2=1** masc=0, fem=1
- **label drop s**
- **label define sexo 0 "masculino" 1 "feminino",**
- **label values sexo sexo**

Baixar o programa escrito por usuário do Stata, que completa o commando logistic:

- **findit prvalue (clicar em st0094)**

Os comandos **cc** e **cs** usam

Exposed= 1 (terapia alternada) **Unexposed**= 0 (terapia sequencial)
Cases=1 (resultado piora) **Controls (Noncases)**= 0 (resultado melhora)

```
. cc  resultado terapia,woolf
```

	terapia		Proportion	
	Exposed	Unexposed	Total	Exposed
Cases	104	89	193	0.5389
Controls	44	62	106	0.4151
Total	148	151	299	0.4950
	Point estimate		[95% Conf. Interval]	
Odds ratio	1.646578		1.019976	2.658121 (Woolf)
Attr. frac. ex.	.3926799		.0195849	.6237944 (Woolf)
Attr. frac. pop	.2115995			

chi2(1) = 4.19 Pr>chi2 = 0.0406

```
. cs  resultado terapia
```

	terapia		
	Exposed	Unexposed	Total
Cases	104	89	193
Noncases	44	62	106
Total	148	151	299
Risk	.7027027	.589404	.6454849
	Point estimate		[95% Conf. Interval]
Risk difference	.1132987		.0056922 .2209052
Risk ratio	1.192226		1.006421 1.412334
Attr. frac. ex.	.1612328		.0063802 .2919521
Attr. frac. pop	.0868819		

chi2(1) = 4.19 Pr>chi2 = 0.0406

- **tab resultado terapia,row chi**

```
+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

           |        terapia
 resultado |         0          1 |     Total
-----------+----------------------+----------
         0 |        62         44 |       106
           |     41.06      29.73 |     35.45
-----------+----------------------+----------
         1 |        89        104 |       193
           |     58.94      70.27 |     64.55
-----------+----------------------+----------
     Total |       151        148 |       299
           |    100.00     100.00 |    100.00

          Pearson chi2(1) =   4.1927   Pr = 0.041
```

- **table terapia,c(freq mean resultado)**

```
-------------------------------------------
  terapia |          Freq.  mean(result~o)
----------+--------------------------------
      seq |             151         .589404
      alt |             148        .7027027
          |
    Total |             299        .6454849
-------------------------------------------
```

O comando logit considera caso o valor 1 e controle o valor 0, portanto resultado mau =1 = caso (piora) e resultado bom =0 = controle (melhora).

- **logit resultado terapia**

```
Iteration 0:   log likelihood = -194.40888
Iteration 1:   log likelihood = -192.30753
Iteration 2:   log likelihood = -192.30471

Logit estimates                                   Number of obs   =        299
                                                  LR chi2(1)      =       4.21
                                                  Prob > chi2     =     0.0402
Log likelihood = -192.30471                       Pseudo R2       =     0.0108

------------------------------------------------------------------------------
   resultado |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     terapia |   .4986993   .2443508     2.04   0.041     .0197805     .977618
       _cons |    .361502   .1654236     2.19   0.029     .0372777    .6857263
------------------------------------------------------------------------------
```

Ou

Statistics → Binary outcomes ⟶ Logistic regression

O algoritmo precisa de 3 iterações para convergir. O coeficiente de terapia representa a diferença no *log odds* de resultado pior entre as terapias alternada e seqüencial. O valor >0 indica que a terapia alternada é pior que a terapia sequencial. O valor de *p* associado à estatística *z* do teste de *Wald* é 0,041. A estatística *z* é igual ao coeficiente dividido pelo erro padrão. Este valor de *p* é assintoticamente igual ao valor de *p* derivado do teste da razão de verossimilhança entre o modelo incluindo somente a constante e o modelo incluindo a variável terapia (chi2(1gl)=4,21). -2 vezes o logaritmo da razão de verossimilhança é igual a 4,21 com distribuição aproximadamente qui quadrado, com 1 grau de liberdade, com valor *p*= 0,040.

- **logistic resultado terapia**

```
Logit estimates                                   Number of obs   =        299
                                                  LR chi2(1)      =       4.21
                                                  Prob > chi2     =     0.0402
Log likelihood = -192.30471                       Pseudo R2       =     0.0108

------------------------------------------------------------------------------
   resultado | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     terapia |   1.646578   .4023427     2.04   0.041     1.019977    2.658117
------------------------------------------------------------------------------
```

exp(0,4986993)= 1,646578

Ou

Statistics → Binary outcomes ⟶ Logistic regression (repporting odds ration)

- **predict pr, pr**
- **table terapia,c(freq mean pr) col row scol**

```
----------------------------------
  terapia |      Freq.    mean(pr)
----------+-----------------------
       seq |         151     .589404
       alt |         148    .7027027
           |
     Total |         299     .645485
```

Obs: os resultados modelados e não modelados são iguais, pois só há uma variável explicativa, dicotômica.

- **`prvalue,x(terapia=0)`**

logit: Predictions for resultado

Confidence intervals by delta method

		95% Conf. Interval	
Pr(y=1|x):	**0.5894**	**[0.5109,**	**0.6679]**
Pr(y=0|x):	**0.4106**	**[0.3321,**	**0.4891]**

comparar com "table terapia"

terapia

x= 0

- **`prvalue,x(terapia=1)`**

logit: Predictions for resultado

Confidence intervals by delta method

		95% Conf. Interval	
Pr(y=1|x):	**0.7027**	**[0.6291,**	**0.7763]**
Pr(y=0|x):	**0.2973**	**[0.2237,**	**0.3709]**

comparar com "table terapia"

terapia

x= 1

- **logistic resultado terapia sexo**

```
Logit estimates                                   Number of obs   =        299
                                                  LR chi2(2)      =       7.55
                                                  Prob > chi2     =     0.0229
Log likelihood = -190.63171                       Pseudo R2       =     0.0194

------------------------------------------------------------------------------
   resultado | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     terapia |   1.652355   .4059667     2.04   0.041     1.020873    2.674452
        sexo |   1.923819   .7146486     1.76   0.078      .928892    3.984405
------------------------------------------------------------------------------
```

- **predict sexter,pr**

- **table terapia sexo,c(freq mean sexter) col row scol (modelado)**

terapia	sexo masculino	feminino	Total
seq	128	23	151
	.5667254	.7156155	.589404
alt	125	23	148
	.6836733	.8061237	.7027027
Total	253	46	299
	.6245059	.7608696	.645485

- **prtab terapia sexo**

```
logistic: Predicted probabilities of posi-
tive outcome for resultado

------------------------------------
           |             sexo
  terapia  | masculino(0)   feminino(1)
----------+-------------------------
         0 |    0.5667        0.7156
         1 |    0.6837        0.8061
------------------------------------

      terapia       sexo
x=   .49498328   .15384615
```

- **prvalue,x(terapia=1 sexo=1)**

logistic: Predictions for resultado

Confidence intervals by delta method

		95% Conf. Interval	
Pr(y=1|x):	**0.8061**	**[0.6914,**	**0.9208]**
Pr(y=0|x):	**0.1939**	**[0.0792,**	**0.3086]**

	terapia	sexo
x=	**1**	**1**

- **table terapia sexo,c(freq mean resultado) col row scol**
  (não modelado)

terapia	sexo masculino (0)	feminino (1)	Total
(0) seq	128	23	151
	.5703125	.6956522	.589404
(1) alt	125	23	148
	.68	.8260869	.7027027
Total	253	46	299
	.6245059	.7608696	.6454849

- poisson resultado terapia sexo, robust irr

```
. poisson  resultado terapia sexo,robust irr

Iteration 0:   log pseudolikelihood = -276.21258
Iteration 1:   log pseudolikelihood = -276.21258

Poisson regression                                Number of obs   =        299
                                                  Wald chi2(2)    =       8.56
                                                  Prob > chi2     =     0.0138
Log pseudolikelihood = -276.21258                 Pseudo R2       =     0.0046
```

resultado	IRR	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
terapia	1.191453	.1026371	2.03	0.042	1.006353	1.410598
sexo	1.217092	.1157873	2.07	0.039	1.010055	1.466567

5.3 - Regressão logística (caso-controle; razão de forças de morbidade)

DADOS DO ESTUDO DE CÂNCER DO ESÔFAGO (SOUZA ET ALII, 1980)

Abrir o banco cânceresôfago.dta

D = CÂNCER DO ESÔFAGO (CASO: 1/CONTROLE: 0)
E = FATOR DE RISCO: FUMAR (SIM: 1/NÃO: 0)
C = VARIÁVEL DE CONFUSÃO: BEBER (SIM: 1/NÃO: 0)

	CÂNCER		
FUMAR	CASO(1)	CONTROLE(0)	TOTAL
SIM(1)	85	113	198
NÃO(0)	13	80	93
TOTAL	98	193	291

$$OR = \frac{85 x 80}{113 x 13} = 4,6$$

$$x^2 = \frac{(85 - 66,7)^2}{14,18} = 23,62$$

BEBER= SIM (1)

	CÂNCER		
FUMAR	CA(1)	CO(0)	TOTAL
SIM(1)	79	70	149
NÃO(0)	5	16	21
TOTAL	84	86	170

$$OR = \frac{79 x 16}{70 x 5} = 3,6$$

$$x^2 = \frac{(79 - 73,6)^2}{149 x 21 x 84 x 86 / 170^2 x 169} = 6,3$$

BEBER=NÃO(0)

	CÂNCER		
FUMAR	CA(1)	CO(0)	TOTAL
SIM(1)	6	43	49
NÃO(0)	8	64	72
TOTAL	14	107	121

$$OR = \frac{6 x 64}{43 x 8} = 1,1$$

$$x^2 = \frac{(6 - 5,7)^2}{49 x 72 x 14 x 107 / 121^2 x 120} = 0,03$$

- ***list***

```
. list
```

	beber	fumar	caco	pop
1.	0	0	0	64
2.	1	0	0	16
3.	0	0	1	8
4.	0	1	1	6
5.	0	1	0	43
6.	1	1	1	79
7.	1	1	0	70
8.	1	0	1	5

- **logistic caco beber fumar [freq=pop]**

Logistic regression **Number of obs = 291**
LR chi2(2) = 53.78
Prob > chi2 = 0.0000
Log likelihood = -159.02123 **Pseudo R2 = 0.1446**

caco	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
beber	5.486778	1.916609	4.87	0.000	2.766819	10.88063
fumar	2.142026	.8002371	2.04	0.041	1.029972	4.454759

- **logit caco beber fumar [freq=pop]**

caco	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
beber	1.702341	.349314	4.87	0.000	1.017698	2.386984
fumar	.7617523	.3735888	2.04	0.041	.0295316	1.493973
_cons	-2.396791	.3521402	-6.81	0.000	-3.086973	-1.706609

5.4- Regressão logística condicional (dados pareados)

Abrir o banco de dados c:\cursostata\cacopareado.dta

- **use "C:\HEPStata\VERÃO\2013\cacopareado.dta"**

***caco: 0= peso normal, controle; 1= peso baixo, caso**

***fumar: 0= mãe não fuma; 1= mãe fuma**

- **list in 1/6**

```
     +--------------------------+
     | id   caco   fumar   idade |
     |--------------------------|
  1. |  1      0       0      14 |
  2. |  1      1       1      14 |
  3. |  2      0       0      18 |
  4. |  2      1       0      18 |
  5. |  3      0       0      16 |
     |--------------------------|
  6. |  3      1       0      16 |
     +--------------------------+
```

- **list in 35/40, clean**

```
      id   caco   fumar   idade
35.   18      0       1      15
36.   18      1       0      15
37.   19      0       0      18
38.   19      1       1      18
39.   20      0       0      21
40.   20      1       1      21
```

- **foreach num of numlist 1 2 9/10 19 20{**

2. display `num'

3. tab caco fumar if id==`num'

4. }

```
1

           |        fumar
      caco |        0          1 |     Total
-----------+----------------------+----------
         0 |        1          0 |         1
         1 |        0          1 |         1
-----------+----------------------+----------
     Total |        1          1 |         2
2

           |   fumar
      caco |        0 |     Total
-----------+----------+----------
         0 |        1 |         1
         1 |        1 |         1
-----------+----------+----------
     Total |        2 |         2
9

           |        fumar
      caco |        0          1 |     Total
-----------+----------------------+----------
         0 |        0          1 |         1
         1 |        1          0 |         1
-----------+----------------------+----------
     Total |        1          1 |         2
10

           |   fumar
      caco |        1 |     Total
-----------+----------+----------
         0 |        1 |         1
         1 |        1 |         1
-----------+----------+----------
     Total |        2 |         2
```

```
19

           |        fumar
      caco |        0          1 |     Total
-----------+----------------------+----------
         0 |        1          0 |         1
         1 |        0          1 |         1
-----------+----------------------+----------
     Total |        1          1 |         2

20

           |        fumar
      caco |        0          1 |     Total
-----------+----------------------+----------
         0 |        1          0 |         1
         1 |        0          1 |         1
-----------+----------------------+----------
     Total |        1          1 |         2
```

- **forvalues i=1(1)6{**

  2. display `i'

  3. tab caco fumar if id==`i'

  4. }

```
1
           |        fumar
      caco |        0          1 |     Total
-----------+----------------------+----------
         0 |        1          0 |         1
         1 |        0          1 |         1
-----------+----------------------+----------
     Total |        1          1 |         2

2
           |   fumar
      caco |        0 |     Total
-----------+----------+----------
         0 |        1 |         1
         1 |        1 |         1
```

```
-----------+-----------+----------
     Total |         2 |         2
```

```
3
```

```
           |   fumar
      caco |         0 |     Total
-----------+-----------+----------
         0 |         1 |         1
         1 |         1 |         1
-----------+-----------+----------
     Total |         2 |         2
```

```
4
```

```
           |         fumar
      caco |         0          1 |     Total
-----------+----------------------+----------
         0 |         1          0 |         1
         1 |         0          1 |         1
-----------+----------------------+----------
     Total |         1          1 |         2
```

```
5
```

```
           |   fumar
      caco |         1 |     Total
-----------+-----------+----------
         0 |         1 |         1
         1 |         1 |         1
-----------+-----------+----------
     Total |         2 |         2
```

```
6
```

```
           |         fumar
      caco |         0          1 |     Total
-----------+----------------------+----------
         0 |         1          0 |         1
         1 |         0          1 |         1
-----------+----------------------+----------
     Total |         1          1 |         2
```

- **clogit caco fumar idade,group(id)**

```
note: idade omitted because of no within-group variance.

Iteration 0:   log likelihood = -11.453857
Iteration 1:   log likelihood = -11.453857

Conditional (fixed-effects) logistic regression   Number of obs   =         40
                                                  LR chi2(1)      =       4.82
                                                  Prob > chi2     =     0.0282
Log likelihood = -11.453857                       Pseudo R2       =     0.1738

------------------------------------------------------------------------------
       caco |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
      fumar |   1.504077    .781736     1.92   0.054    -.0280969    3.036252
------------------------------------------------------------------------------
```

- **clogit,or**

```
Conditional (fixed-effects) logistic regression   Number of obs   =         40
                                                  LR chi2(1)      =       4.82
                                                  Prob > chi2     =     0.0282
Log likelihood = -11.453857                       Pseudo R2       =     0.1738

------------------------------------------------------------------------------
       caco | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
      fumar |        4.5   3.517812     1.92   0.054     .9722941    20.82703
------------------------------------------------------------------------------
```

- **reshape wide fumar,i(id) j(caco)** --------->reshape

  (note: j = 0 1) wide varquepodemudar (00 01 10 11 no par),i(id) j(varpareada) só é 00 ou 11 no par.

```
Data                               long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                       40   ->      20
Number of variables                   4   ->       4
j variable (2 values)              caco   ->   (dropped)
xij variables:
                                  fumar   ->   fumar0 fumar1
-----------------------------------------------------------------------------
```

- **mcc fumar1 fumar0**

```
                | Controls               |
Cases           |   Exposed   Unexposed  |      Total
----------------+------------------------+-----------
         Exposed |         3           9  |         12
       Unexposed |         2           6  |          8
----------------+------------------------+-----------
           Total |         5          15  |         20

McNemar's chi2(1) =      4.45    Prob > chi2 = 0.0348
Exact McNemar significance probability       = 0.0654

Proportion with factor
        Cases             .6
        Controls         .25     [95% Conf. Interval]
                    ---------     --------------------
        difference       .35      .0134495   .6865505
        ratio            2.4      1.036928   5.554871
        rel. diff.   .4666667      .1501821   .7831512

        odds ratio       4.5      .9314123   42.79972   (exact)
```

- **list in 1/10**

```
     +-----------------------------+
     | id   fumar0   fumar1   idade |
     |-----------------------------|
  1. |  1        0        1      14 |
  2. |  2        0        0      18 |
  3. |  3        0        0      16 |
  4. |  4        0        1      19 |
  5. |  5        1        1      20 |
     |-----------------------------|
  6. |  6        0        1      14 |
  7. |  7        0        0      15 |
  8. |  8        0        0      15 |
  9. |  9        1        0      17 |
 10. | 10        1        1      18 |
     +-----------------------------+
```

- **clogit peso fumanãofuma idade,group(id)**

```
note: multiple positive outcomes within groups encountered.
note: 9 groups (18 obs) dropped because of all positive or
      all negative outcomes.
note: idade omitted because of no within-group variance.

Iteration 0:   log likelihood = -6.5465327
Iteration 1:   log likelihood = -5.2155344
Iteration 2:   log likelihood = -5.2155324
Iteration 3:   log likelihood = -5.2155324

Conditional (fixed-effects) logistic regression   Number of obs   =         22
                                                  LR chi2(1)      =       4.82
                                                  Prob > chi2     =     0.0282
Log likelihood = -5.2155324                       Pseudo R2       =     0.3160

------------------------------------------------------------------------------
        peso |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 fumanãofuma |   1.504077    .781736     1.92   0.054    -.0280969    3.036252
------------------------------------------------------------------------------
```

- **clogit peso fumanãofuma idade,group(id) or**

```
note: multiple positive outcomes within groups encountered.
note: 9 groups (18 obs) dropped because of all positive or
      all negative outcomes.
note: idade omitted because of no within-group variance.

Iteration 0:   log likelihood = -6.5465327
Iteration 1:   log likelihood = -5.2155344
Iteration 2:   log likelihood = -5.2155324
Iteration 3:   log likelihood = -5.2155324

Conditional (fixed-effects) logistic regression   Number of obs   =         22
                                                  LR chi2(1)      =       4.82
                                                  Prob > chi2     =     0.0282
Log likelihood = -5.2155324                       Pseudo R2       =     0.3160

------------------------------------------------------------------------------
        peso | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 fumanãofuma |        4.5   3.517812     1.92   0.054     .9722941    20.82703
------------------------------------------------------------------------------
```

- **reshape wide peso,i(id) j(fumanãofuma)**

```
(note: j = 0 1)      ---->peso pode variar no par (00 01 10 11); fumanãofu
                          ma  é sempre 00 ou 11

 Data                               long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                        40   ->      20
Number of variables                    4   ->       4
j variable (2 values)       fumanãofuma   ->   (dropped)
xij variables:
                                    peso   ->   peso0 peso1
-----------------------------------------------------------------------------
```

- **mcc peso1 peso0**

```
                | Controls              |
Cases           |   Exposed   Unexposed  |      Total
----------------+-----------------------+-----------
         Exposed |         3           9  |         12
       Unexposed |         2           6  |          8
----------------+-----------------------+-----------
           Total |         5          15  |         20

McNemar's chi2(1) =      4.45    Prob > chi2 = 0.0348
Exact McNemar significance probability       = 0.0654

Proportion with factor
        Cases             .6
        Controls         .25     [95% Conf. Interval]
                   ---------     --------------------
        difference       .35      .0134495   .6865505
        ratio            2.4      1.036928   5.554871
        rel. diff.  .4666667      .1501821   .7831512

        odds ratio       4.5      .9314123   42.79972   (exact)
```

5.5- Regressão logística multinomial (politômica)

ADAPTADO DO TRABALHO DE AMANDA APARECIDA SILVA ET ALII, 2010)

Abrir o banco amandamlogitinic.dta

- **tab formapreferencialdeestudo matériadoexercício,col**

```
formapreferenc |            matériadoexercício
   ialdeestudo |   desenho  matemátic     música  português |     Total
---------------+--------------------------------------------+----------
  1_emconjunto |       112         97         69         90 |       368
               |     54.63      61.01      45.10      52.33 |     53.41
---------------+--------------------------------------------+----------
2_isoladamente |        43         31         28         33 |       135
               |     20.98      19.50      18.30      19.19 |     19.59
---------------+--------------------------------------------+----------
 3_indiferente |        50         31         56         49 |       186
               |     24.39      19.50      36.60      28.49 |     27.00
---------------+--------------------------------------------+----------
         Total |       205        159        153        172 |       689
               |    100.00     100.00     100.00     100.00 |    100.00
```

- **tab formapreferencialdeestudonum matériadoexercício,col**

```
formaprefe |
rencialdee |            matériadoexercício
  studonum |   desenho  matemátic     música  português |     Total
-----------+--------------------------------------------+----------
         1 |       112         97         69         90 |       368
           |     54.63      61.01      45.10      52.33 |     53.41
-----------+--------------------------------------------+----------
         2 |        43         31         28         33 |       135
           |     20.98      19.50      18.30      19.19 |     19.59
-----------+--------------------------------------------+----------
         3 |        50         31         56         49 |       186
           |     24.39      19.50      36.60      28.49 |     27.00
-----------+--------------------------------------------+----------
     Total |       205        159        153        172 |       689
           |    100.00     100.00     100.00     100.00 |    100.00
```

.

- **mlogit formapreferencialdeestudonum matéri-adoexercício2 matériadoexercício3 matériado-exercício4 ,b(1) rrr**

```
Iteration 0:   log likelihood = -694.40574
Iteration 1:   log likelihood = -687.85205
Iteration 2:   log likelihood = -687.79849
Iteration 3:   log likelihood = -687.79848

Multinomial logistic regression                 Number of obs   =        689
                                                LR chi2(6)      =      13.21
                                                Prob > chi2     =     0.0398
Log likelihood = -687.79848                     Pseudo R2       =     0.0095

------------------------------------------------------------------------------
formaprefe~m |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
2            |
matériadoe~2 |   .8324143   .2275884    -0.67   0.502      .487096     1.42254
matériadoe~3 |    1.05696   .3033888     0.19   0.847     .6021857    1.855182
matériadoe~4 |   .9550388   .2590929    -0.17   0.865     .5611747    1.625339
-------------+----------------------------------------------------------------
3            |
matériadoe~2 |   .7158763   .1914162    -1.25   0.211     .4238748    1.209034
matériadoe~3 |   1.817971   .4500294     2.41   0.016     1.119119    2.953232
matériadoe~4 |   1.219556   .2998417     0.81   0.419     .7532237      1.9746
------------------------------------------------------------------------------
(formapreferencialdeestudonum==1 is the base outcome)
```

- **gen formapreferencialdeestudonum12=0 if formapreferencialdeestudonum==1**
- **replace formapreferencialdeestudonum12=1 if formapreferencialdeestudonum==2**
- **tab formapreferencialdeestudonum12**

```
formaprefer |
encialdeest |
   udonum12 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        368       73.16       73.16
          1 |        135       26.84      100.00
------------+-----------------------------------
      Total |        503      100.00
```

- **logistic formapreferencialdeestudonum12 matériadoexercício2 matériadoexercício3 matériadoexercício4**

```
Logistic regression                               Number of obs   =        503
                                                  LR chi2(3)      =       0.72
                                                  Prob > chi2     =     0.8682
Log likelihood = -292.20967                       Pseudo R2       =     0.0012

------------------------------------------------------------------------------
formapref~12 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
matériadoe~2 |   .8324143   .2275882    -0.67   0.502     .4870962     1.42254
matériadoe~3 |    1.05696   .3033887     0.19   0.847     .6021857    1.855182
matériadoe~4 |   .9550388   .2590929    -0.17   0.865     .5611747    1.625339
------------------------------------------------------------------------------
```

- **gen formapreferencialdeestudonum13=0 if formapreferencialdeestudonum==1**

```
(321 missing values generated)
```

- **replace formapreferencialdeestudonum13=1 if formapreferencialdeestudonum==3**

```
(186 real changes made)
```

- **tab formapreferencialdeestudonum13**

```
formaprefer |
encialdeest |
   udonum13 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        368       66.43       66.43
          1 |        186       33.57      100.00
------------+-----------------------------------
      Total |        554      100.00
```

- **logistic formapreferencialdeestudonum13 matériadoexercício2 matériadoexercício3 matériadoexercício4**

```
Logistic regression                               Number of obs   =        554
                                                  LR chi2(3)      =      12.79
                                                  Prob > chi2     =     0.0051
Log likelihood = -347.15287                       Pseudo R2       =     0.0181

------------------------------------------------------------------------------
formapref~13 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
matériadoe~2 |   .7158763   .1914162    -1.25   0.211     .4238748    1.209034
matériadoe~3 |   1.817971   .4500294     2.41   0.016     1.119119    2.953232
matériadoe~4 |   1.219556   .2998417     0.81   0.419     .7532237      1.9746
------------------------------------------------------------------------------
```

5.6- Regressão logística ordinal

ADAPTADO DO TRABALHO DE AMANDA APARECIDA SILVA ET ALII, 2010)

Abrir o banco amandagologit2inic.dta

- **gen níveldestress23=11 if níveldestress==1**

(321 missing values generated)

- **replace níveldestress23=23 if níveldestress==2| níveldestress==3**

(321 real changes made)

- **gen níveldestress12=12 if níveldestress==1| níveldestress==2**

(186 missing values generated)

- **replace níveldestress12=33 if níveldestress==3**

(186 real changes made)

- **tab níveldestress graudedemanda if graudedemanda==1| graudedemanda==4,col chi**

```
+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

níveldestr |     graudedemanda
       ess |         1          4 |     Total
-----------+----------------------+----------
         1 |        97         69 |       166
           |     61.01      45.10 |     53.21
-----------+----------------------+----------
         2 |        31         28 |        59
           |     19.50      18.30 |     18.91
-----------+----------------------+----------
         3 |        31         56 |        87
           |     19.50      36.60 |     27.88
-----------+----------------------+----------
     Total |       159        153 |       312
           |    100.00     100.00 |    100.00

          Pearson chi2(2) =  11.9484   Pr = 0.003
```

- **dis 56*31/28/31**
  OR= 2

- **dis 28*97/69/31**
  OR= 1.2697522

- **tab níveldestress23 graudedemanda if graudedemanda==1| graudedemanda==4,col chi**

```
+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

níveldestr |     graudedemanda
     ess23 |         1          4 |     Total
-----------+----------------------+----------
        11 |        97         69 |       166
           |     61.01      45.10 |     53.21
-----------+----------------------+----------
        23 |        62         84 |       146
           |     38.99      54.90 |     46.79
-----------+----------------------+----------
     Total |       159        153 |       312
           |    100.00     100.00 |    100.00

          Pearson chi2(1) =   7.9255   Pr = 0.005
```

- **dis 84*97/69/62**

OR= 1.9046283

- **tab níveldestress12 graudedemanda if graudedemanda==1| graudedemanda==4,col chi**

```
+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

níveldestr |     graudedemanda
     ess12 |         1          4 |     Total
-----------+----------------------+----------
        12 |       128         97 |       225
           |     80.50      63.40 |     72.12
-----------+----------------------+----------
        33 |        31         56 |        87
           |     19.50      36.60 |     27.88
-----------+----------------------+----------
     Total |       159        153 |       312
           |    100.00     100.00 |    100.00

          Pearson chi2(1) =  11.3438   Pr = 0.001
```

- **dis 56*128/97/31**

OR= 2.3837712

- **ssc install gologit2**

    - **use "C:\HEPStata\DIFUSÃO\StatadifusãoVIII_nov2012\amandagologit2inic.dta", replace**

    - **tab níveldestress graudedemanda,col chi**

```
+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

níveldestr |                graudedemanda
       ess |         1          2          3          4 |     Total
-----------+--------------------------------------------+----------
         1 |        97         90        112         69 |       368
           |     61.01      52.33      54.63      45.10 |     53.41
-----------+--------------------------------------------+----------
         2 |        31         33         43         28 |       135
           |     19.50      19.19      20.98      18.30 |     19.59
-----------+--------------------------------------------+----------
         3 |        31         49         50         56 |       186
           |     19.50      28.49      24.39      36.60 |     27.00
-----------+--------------------------------------------+----------
     Total |       159        172        205        153 |       689
           |    100.00     100.00     100.00     100.00 |    100.00

          Pearson chi2(6) =  13.3368   Pr = 0.038
```

OR(2+3 vs 1): 1,43 1,30 1,90
OR(3 vs 2+3): 1,64 1.63 2,38

- **tab níveldestress23 graudedemanda,col chi**

```
+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

níveldestr |                graudedemanda
     ess23 |         1          2          3          4 |     Total
-----------+----------------------------------------------+----------
        11 |        97         90        112         69 |       368
           |     61.01      52.33      54.63      45.10 |     53.41
-----------+----------------------------------------------+----------
        23 |        62         82         93         84 |       321
           |     38.99      47.67      45.37      54.90 |     46.59
-----------+----------------------------------------------+----------
     Total |       159        172        205        153 |       689
           |    100.00     100.00     100.00     100.00 |    100.00

          Pearson chi2(3) =   8.1398   Pr = 0.043
```

- **tab níveldestress12 graudedemanda,col chi**

```
+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

níveldestr |                graudedemanda
     ess12 |         1          2          3          4 |     Total
-----------+----------------------------------------------+----------
        12 |       128        123        155         97 |       503
           |     80.50      71.51      75.61      63.40 |     73.00
-----------+----------------------------------------------+----------
        33 |        31         49         50         56 |       186
           |     19.50      28.49      24.39      36.60 |     27.00
-----------+----------------------------------------------+----------
     Total |       159        172        205        153 |       689
           |    100.00     100.00     100.00     100.00 |    100.00

          Pearson chi2(3) =  12.6004   Pr = 0.006
```

- **tab graudedemanda,gen(graudedemanda)**

```
graudedeman |
          da |      Freq.     Percent        Cum.
-------------+-----------------------------------
           1 |        159       23.08       23.08
           2 |        172       24.96       48.04
           3 |        205       29.75       77.79
           4 |        153       22.21      100.00
-------------+-----------------------------------
       Total |        689      100.00
```

- **gologit2 níveldestress graudedemanda2 graudedemanda3 graudedemanda4,or**
  (valores modelados iguais aos observados)

```
Generalized Ordered Logit Estimates              Number of obs   =        689
                                                 LR chi2(6)      =      13.21
                                                 Prob > chi2     =     0.0398
Log likelihood = -687.79848                      Pseudo R2       =     0.0095

------------------------------------------------------------------------------
níveldestr~s | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
graudedema~2 |   1.425448   .3179246     1.59   0.112     .9206695    2.206983
graudedema~3 |   1.299107   .2789894     1.22   0.223     .8527966    1.978994
graudedema~4 |   1.904628   .4377987     2.80   0.005     1.213815      2.9886
-------------+----------------------------------------------------------------
2            |
graudedema~2 |   1.644899    .430853     1.90   0.057     .9844225    2.748507
graudedema~3 |   1.331946   .3435341     1.11   0.266     .8034254    2.208145
graudedema~4 |   2.383771   .6226938     3.33   0.001     1.428604    3.977565
------------------------------------------------------------------------------
```

- **predict p1 p2 p3**

(option p assumed; predicted probabilities)

- **egen numobs=count (graudedemanda),by(graudedemanda)**
- **gen obsp1=numobs*p1**
- **gen obsp2=numobs*p2**
- **gen obsp3=numobs*p3**

- **replace obsp1=round(obsp1)**

(153 real changes made)

- **replace obsp2=round(obsp2)**

(325 real changes made)

- **replace obsp3=round(obsp3)**

(536 real changes made)

- **table níveldestress graudedemanda ,c(freq mean obsp1 mean obsp2 mean obsp3) row col scol**

níveldestress	graudedemanda 1	2	3	4	Total
1	97	90	112	69	368
	97	90	112	69	94.60326
	31	33	43	28	34.5788
	31	49	50	56	45.87228
2	31	33	43	28	135
	97	90	112	69	94.25926
	31	33	43	28	34.68889
	31	49	50	56	46.63704
3	31	49	50	56	186
	97	90	112	69	90.75806
	31	33	43	28	33.84946
	31	49	50	56	48.37634
Total	159	172	205	153	689
	97	90	112	69	93.49783
	31	33	43	28	34.40348
	31	49	50	56	46.69811

- **gologit2 níveldestress graudedemanda2 graudedemanda3 graudedemanda4,or autofit**

```
------------------------------------------------------------------------------
Testing parallel lines assumption using the .05 level of significance...

Step  1:  Constraints for parallel lines imposed for graudedemanda3 (P Value =
0.90
> 54)
Step  2:  Constraints for parallel lines imposed for graudedemanda2 (P Value =
0.44
> 57)
Step  3:  Constraints for parallel lines imposed for graudedemanda4 (P Value =
0.28
> 26)
Step  4:  All explanatory variables meet the pl assumption

Wald test of parallel lines assumption for the final model:

 ( 1)  [1]graudedemanda3 - [2]graudedemanda3 = 0
 ( 2)  [1]graudedemanda2 - [2]graudedemanda2 = 0
 ( 3)  [1]graudedemanda4 - [2]graudedemanda4 = 0

           chi2(  3) =    1.70
         Prob > chi2 =    0.6375

An insignificant test statistic indicates that the final model
does not violate the proportional odds/ parallel lines assumption

If you re-estimate this exact same model with gologit2, instead
of autofit you can save time by using the parameter

pl(graudedemanda3 graudedemanda2 graudedemanda4)

------------------------------------------------------------------------------

Generalized Ordered Logit Estimates               Number of obs   =        689
                                                  Wald chi2(3)    =      11.45
                                                  Prob > chi2     =     0.0095
Log likelihood = -688.63433                       Pseudo R2       =     0.0083

 ( 1)  [1]graudedemanda3 - [2]graudedemanda3 = 0
 ( 2)  [1]graudedemanda2 - [2]graudedemanda2 = 0
 ( 3)  [1]graudedemanda4 - [2]graudedemanda4 = 0
------------------------------------------------------------------------------
níveldestr~s | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
graudedema~2 |   1.483397   .3179653     1.84   0.066     .9745467    2.257937
graudedema~3 |   1.300796   .2685266     1.27   0.203     .8679478    1.949508
graudedema~4 |   2.063686   .4528829     3.30   0.001     1.342285      3.1728
-------------+----------------------------------------------------------------
2            |
graudedema~2 |   1.483397   .3179653     1.84   0.066     .9745467    2.257937
graudedema~3 |   1.300796   .2685266     1.27   0.203     .8679478    1.949508
graudedema~4 |   2.063686   .4528829     3.30   0.001     1.342285      3.1728
------------------------------------------------------------------------------
```

- **predict pauto1 pauto2 pauto3**
  (option p assumed; predicted probabilities)

- **gen obspauto1=numobs*pauto1**

- **gen obspauto2=numobs*pauto2**

- **gen obspauto3=numobs*pauto3**
- **rename obspauto1 obsmodelpauto1**
- **rename obspauto2 obsmodelpauto2**
- **rename obspauto3 obsmodelpauto3**
- **replace obsmodelpauto1=round(obsmodelpauto1)**
  (689 real changes made)
- **replace obsmodelpauto2=round(obsmodelpauto2)**
  (689 real changes made)
- **replace obsmodelpauto3=round(obsmodelpauto3)**
  (689 real changes made)

- **table níveldestress graudedemanda, c(freq mean obsmodelpauto1 mean obsmode lpauto2 mean obsmodelpauto3)row col scol**

níveldestress	graudedemanda				
	1	2	3	4	Total
1	97	90	112	69	368
	98	89	113	67	94.55163
	28	35	40	33	34.30163
	33	48	52	54	46.38859
2	31	33	43	28	135
	98	89	113	67	94.14815
	28	35	40	33	34.57037
	33	48	52	54	47.07407
3	31	49	50	56	186
	98	89	113	67	90.32796
	28	35	40	33	34.57527
	33	48	52	54	48.38172
Total	159	172	205	153	689
	98	89	113	67	93.33237
	28	35	40	33	34.42816
	33	48	52	54	47.06096

OR(2+3 vs 1): 1,50 1,31 2,09
OR(3 vs 2+3): 1,49 1,30 2,06

- **ologit níveldestress graudedemanda2 graudedemanda3 graudedemanda4,or**

```
Iteration 0:   log likelihood = -694.40574
Iteration 1:   log likelihood = -688.64181
Iteration 2:   log likelihood = -688.63433
Iteration 3:   log likelihood = -688.63433

Ordered logistic regression                     Number of obs   =        689
                                                LR chi2(3)      =      11.54
                                                Prob > chi2     =     0.0091
Log likelihood = -688.63433                     Pseudo R2       =     0.0083

------------------------------------------------------------------------------
níveldestr~s | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
graudedema~2 |   1.483397   .3179653     1.84   0.066     .9745467    2.257937
graudedema~3 |   1.300796   .2685266     1.27   0.203     .8679478    1.949508
graudedema~4 |   2.063686   .4528829     3.30   0.001     1.342285      3.1728
-------------+----------------------------------------------------------------
       /cut1 |   .4722798   .1581763                      .16226    .7822996
       /cut2 |   1.342695   .1655409                     1.01824    1.667149
------------------------------------------------------------------------------
```

Tabelas usando valores ajustados pelo modelo proporcional -lembrar arredondamentos para números inteiros-

níveldestr ess	graudedemanda 1	4	Total
1	98	67	165
	61.64	43.79	
2	28	33	61
	17.61	21.57	
3	33	53	86
	20.75	34.64	
Total	159	153	312
	100.00	100.00	

níveldestr ess23	graudedemanda 1	4	Total
11	98	67	165
23	61	86	147
Total	159	153	312
	100.00	100.00	

- **dis 86*98/67/61**

OR= 2.062

níveldestr ess12	graudedemanda 1	4	Total
12	126	100	226
33	33	53	86
Total	159	153	312
	100.00	100.00	

dis 53*126/100/33

OR= 2.024

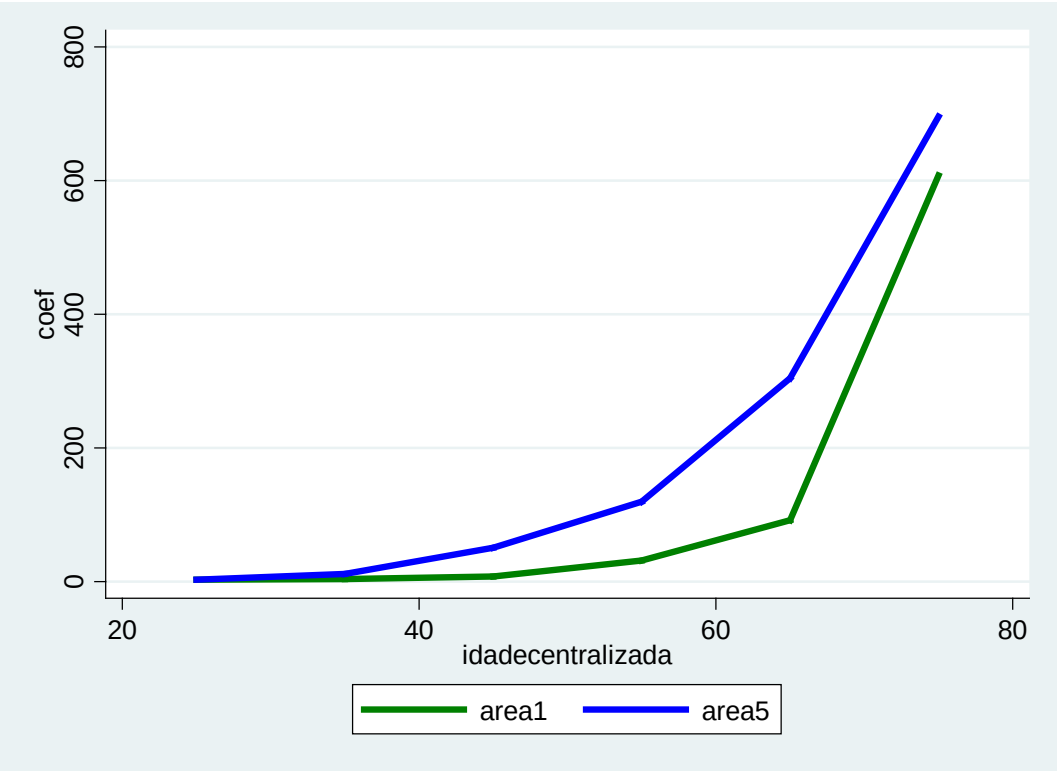
5.7.1 - Regressão de Poisson

Dados parciais de pesquisa de FARIAS, N.;
SOUZA, J.M.P.; LAURENTI, R.; ALENCAR, S.M.

Abrir o arquivo c:/cursostata/norma.dta

Causa de óbito estudada: DCB, doença cérebro-vascular (derrame)
Período de observação: 2003 a 2005
Sexo: 1= M, 0= F
Área: 1= melhor condição sócio-econômica, 5= pior

- **line coef idadecentralizada if sexon==1 & area==1,sort lwidth(thick)lcolor(green)||line coef idadecentralizada if sexon==1 & area==5,sort lcolor(blue)lwidth(thick)**



- **tab area, gen(area)**
- **poisson obito idadecentralizada area2 if sexon==1,exp(pop_3)**

```
Iteration 0:   log likelihood = -128.92382
Iteration 3:   log likelihood = -123.43135
Poisson regression                                Number of obs   =         12
                                                  LR chi2(2)      =    6433.15
                                                  Prob > chi2     =     0.0000
Log likelihood = -123.43135                       Pseudo R2       =     0.9630
------------------------------------------------------------------------------
       obito |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
idadecentr~a |   .1027784   .0014997    68.53   0.000      .099839    .1057179
       area2 |   .6396541   .0519999    12.30   0.000      .537736    .7415721
       _cons |  -13.17455   .1097276  -120.07   0.000    -13.38962   -12.95949
       pop_3 | (exposure)
------------------------------------------------------------------------------
```

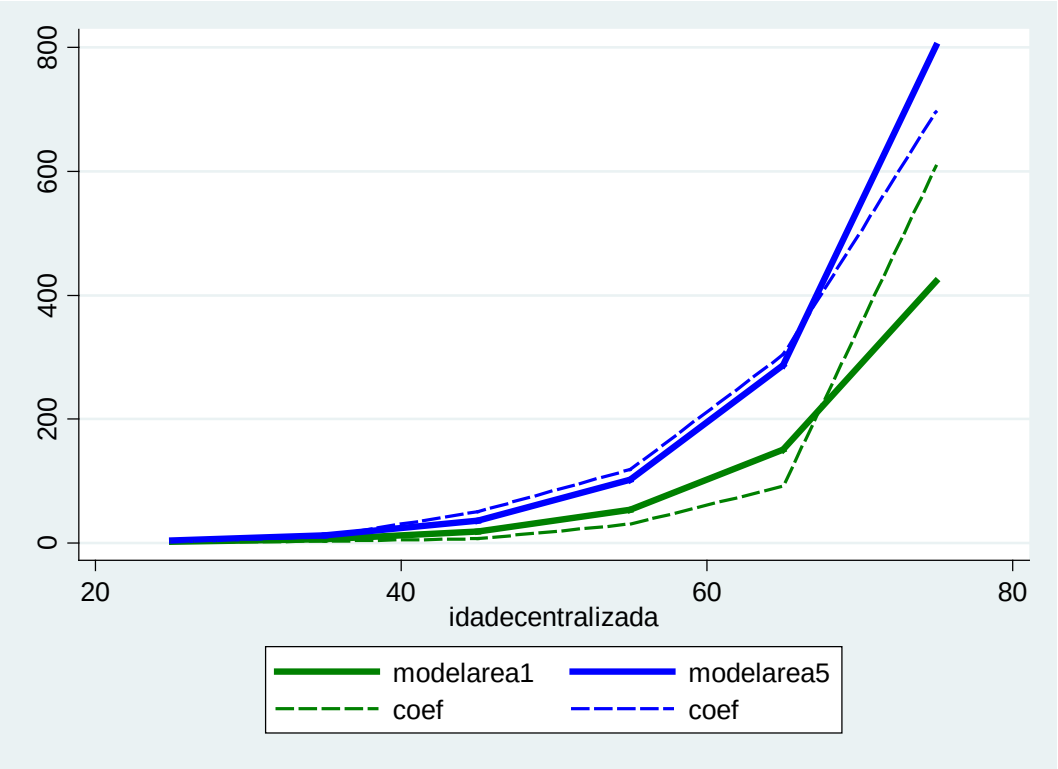
- **predict xb,xb**
- **predict n, n**
- **predict ir,ir**
- **gen coef100000=ir*100000**

- **poisson obito idadecentralizada area2 if sexon==1,exp(pop_3) irr**

```
Iteration 0:   log likelihood = -128.92382
Iteration 3:   log likelihood = -123.43135
Poisson regression                                Number of obs   =         12
                                                  LR chi2(2)      =    6433.15
                                                  Prob > chi2     =     0.0000
Log likelihood = -123.43135                       Pseudo R2       =     0.9630
------------------------------------------------------------------------------
       obito |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
idadecentr~a |   1.108246   .0016621    68.53   0.000     1.104993    1.111508
       area2 |   1.895825   .0985828    12.30   0.000     1.712126    2.099233
       pop_3 | (exposure)
------------------------------------------------------------------------------
```

predict xb,xb = ln(n)
predict n, n = número estimado de óbitos= exp(xb)
predict ir,ir = n/pop_3= coeficiente estimado
gen coef100000=ir*100000 = coeficiente estimado, base 100.000
gen coeflog10=10^(_constant+coef*idadecentralizada+coef*area2)= coeficiente já na base 100.000

- **line coef100000 idadecentralizada if sexon==1 &area==1, sort lwidth(thick)lcolor(green)||line coef100000 idadecentralizada if sexon==1 &area==5, sort lwidth(thick)lcolor(blue)||line coef idadecentralizada if sexon==1 &area==1, sort lcolor(green) lwidth(medthick) lpattern(dash)||line coef idadecentralizada if sexon==1 &area==5, sort lwidth(medthick) lpattern(dash) lcolor(blue)||,legend(cols(2) label(1 "modelarea1") label(2 "modelarea5"))**



Na area 5, na idade de 25 anos, o coeficiente é 4,6999641óbitos/100000anos.

Aos 35 anos o coeficiente será 4,6999641/100000x(1,108246^10)=

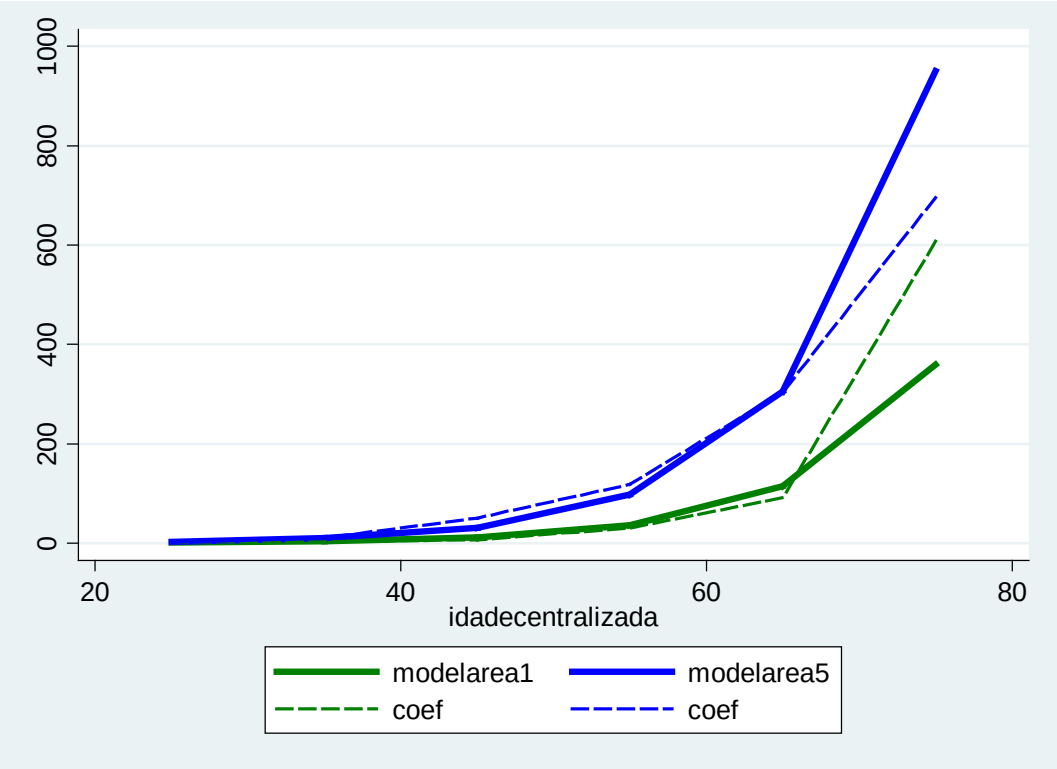
4,6999641/100000x2,794871= 13,13572óbitos/100000anos.

- **reg log10coef idadecentralizada area2 if sexon==1**

```
. reg   log10coef idadecentralizada  area2 if sexon==1 &periodo01==1

      Source |       SS       df       MS              Number of obs =      12
-------------+------------------------------           F(  2,     9) =  135.93
       Model |  9.06771246     2  4.53385623            Prob > F      =  0.0000
    Residual |  .300182195     9  .033353577            R-squared     =  0.9680
-------------+------------------------------           Adj R-squared =  0.9608
       Total |  9.36789466    11  .851626787            Root MSE      =  .18263

------------------------------------------------------------------------------
   log10coef |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
idadecentr~a |   .0493735    .003087    15.99   0.000     .0423902    .0563568
       area2 |   .4225373   .1054413     4.01   0.003     .1840126     .661062
       _cons |  -1.147436   .1714144    -6.69   0.000    -1.535203    -.75967
------------------------------------------------------------------------------
```

- **gen coeflog10=10^(-1.147436+.0493735* idadecentraliza-da+.4225373*area2) if sexon==1**

- **line coeflog10 idadecentralizada if sexon==1 &area==1, sort lwidth(thick)lcolor(green)||line coeflog10 idadecentralizada if sexon==1 &area==5, sort lwidth(thick)lcolor(blue)||line coef idadecentralizada if sexon==1 &area==1, sort lcolor(green) lwidth(medthick)lpattern(dash)||line coef idadecentralizada if sexon==1 &area==5, sort lwidth(medthick) lpattern(dash) lcolor(blue)||,legend(cols(2) label(1 "modelarea1") label(2 "modelarea5"))**

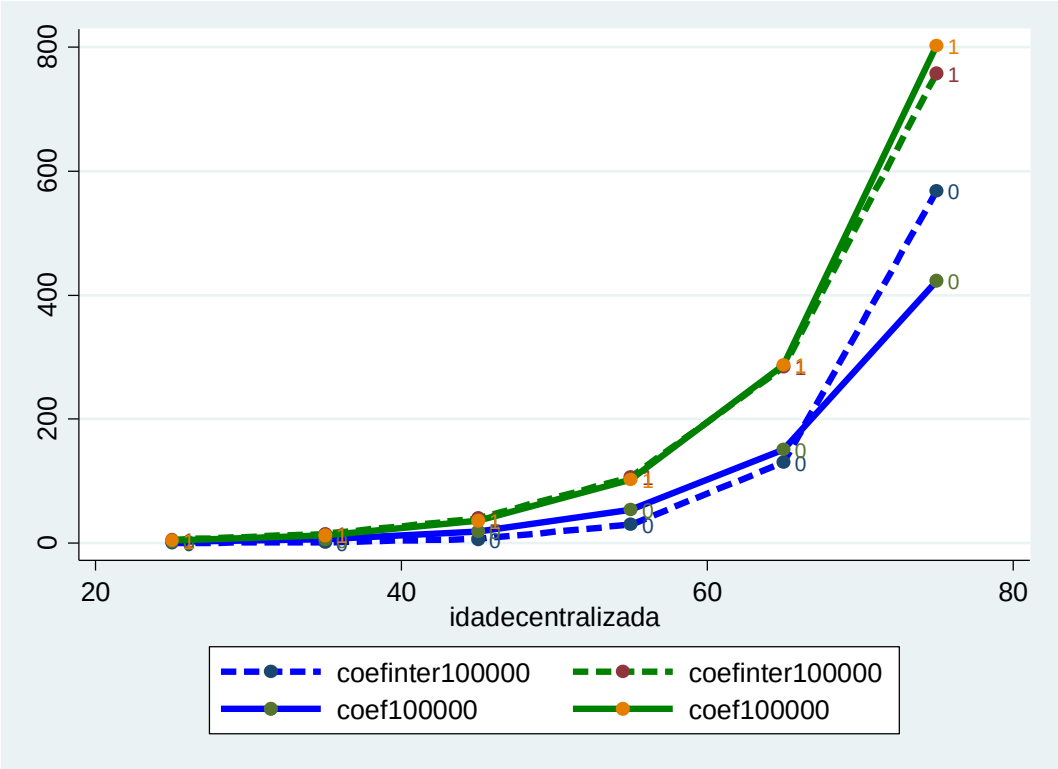


gen inter=idadecentralizada*area2

```
poisson obito idadecentralizada area2  inter if sexon==1, exp( pop_3) irr

Iteration 0:   log likelihood = -210.66325
Iteration 1:   log likelihood = -82.924254
Iteration 2:   log likelihood = -81.658584
Iteration 3:   log likelihood = -81.653747
Iteration 4:   log likelihood = -81.653747

Poisson regression                                Number of obs   =         12
                                                  LR chi2(3)      =    6516.70
                                                  Prob > chi2     =     0.0000
Log likelihood = -81.653747                       Pseudo R2       =     0.9756

------------------------------------------------------------------------------
       obito |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
idadecentra~a |   1.158612   .0067527    25.26   0.000     1.145453    1.171923
       area2 |     53.372   22.63291     9.38   0.000     23.24627    122.5388
       inter |    .952001   .0057435    -8.15   0.000     .9408102    .9633249
       _cons |   9.10e-08   3.75e-08   -39.34   0.000     4.06e-08    2.04e-07
    ln(pop_3) |          1  (exposure)
------------------------------------------------------------------------------
```

- **predict xbinter,xb**
- **predict ninter, n**
- **predict irinter,ir**
- **gen coefinter100000=irinter*100000**

scatter coefinter100000 idadecentralizada if sexon==1 &area==1,connect(L) mlabel(area2) sort lwidth(thick)lcolor(blue)lpattern(dash)||scatter coefinter100000 idade-centralizada if sexon==1 &area==5, connect(L) mlabel(area2)sort lwidth(thick)lcolor(green)lpattern(dash)||scatter coef100000 idade-centralizada if sexon==1 &area==1, connect(L) mlabel(area2)sort lwidth(thick)lcolor(blue)||scatter coef100000 idadecentralizada if se-xon==1 &area==5, connect(L) sort lwidth(thick)lcolor(green)mlabel(area2)


800
600
400
200
0
20
40
60
80
idadecentralizada
coefinter100000
coefinter100000
coef100000
coef100000

5.7.2 - Regressão de Poisson

Dados da dissertação de mestrado Sonia Buongermino de Souza (1987)

- **use c:/cursostata/soniafebre.dta**

- **list leite dias status id tabela leitetabela diastabela falhastabela in 1/5, clean**

	leite	dias	status	id	tabela	leitet~a	diasta~a	falhas~a
1.	0	75	0	1	.	.	.	.
2.	0	2	0	2	.	.	.	.
3.	0	61	0	3	.	.	.	.
4.	0	19	0	4	.	.	.	.
5.	0	18	1	5	.	.	.	.

- **list leite dias status id tabela leitetabela diastabela falhastabela in 75/80, clean**

	leite	dias	status	id	tabela	leitet~a	diasta~a	falhas~a
75.	1	47	0	75	.	.	.	.
76.	1	33	1	76	.	.	.	.
77.	1	13	1	77	.	.	.	.
78.	1	32	1	78	.	.	.	.
79.	.	.	.	.	1	1	1214	17
80.	.	.	.	.	1	0	1438	9

Para a segunda abordagem

- **table leite ,contents(freq sum status sum dias)**

leite	Freq.	sum(status)	sum(dias)
0	44	9	1438
1	34	17	1214

- **poisson status leite,exp(dias) irr**

```
Iteration 0:   log likelihood = -72.361595
Iteration 1:   log likelihood = -72.361581
Iteration 2:   log likelihood = -72.361581

Poisson regression                      Number of obs   =         78
                                        LR chi2(1)      =       4.04
                                        Prob > chi2     =     0.0444
Log likelihood = -72.361581             Pseudo R2       =     0.0272

------------------------------------------------------------------------------
      status |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       leite |   2.237415   .9223326     1.95   0.051     .9973663    5.019247
       dias  | (exposure)
------------------------------------------------------------------------------
```

- **predict ir,ir**

```
(2 missing values generated)
```

- **table leite ,con(freq sum status sum dias mean ir)**

```
-------------------------------------------------------------
      leite |      Freq.  sum(status)   sum(dias)    mean(ir)
------------+------------------------------------------------
          0 |        44           9          1438    .0062587
          1 |        34          17          1214    .0140033
-------------------------------------------------------------
```

Segunda abordagem

- **table leitetabela,contents(freq sum falhastabela sum diastabela)**

```
-----------------------------------------------------
leitetabe |
la        |      Freq.  sum(falhas~a)  sum(diasta~a)
----------+------------------------------------------
        0 |         1           9          1438
        1 |         1          17          1214
-----------------------------------------------------
```

- **poisson falhastabela leitetabela,exp(diastabela) irr**

```
Iteration 0:   log likelihood = -4.3672529
Iteration 1:   log likelihood = -4.3672529

Poisson regression                      Number of obs   =          2
                                        LR chi2(1)      =       4.04
                                        Prob > chi2     =     0.0444
Log likelihood = -4.3672529             Pseudo R2       =     0.3164

------------------------------------------------------------------------------
falhastabela |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  leitetabela |   2.237415   .9223326     1.95   0.051     .9973663    5.019246
  diastabela | (exposure)
------------------------------------------------------------------------------
```

- **predict irtabela,ir**

```
(78 missing values generated)
```

- **table leitetabela,contents(freq sum falhastabela sum diastabela mean irtabela)**

```
-------------------------------------------------------------------------
 leitetabe |
    la     |      Freq.  sum(falhas~a)  sum(diasta~a)  mean(irtabela)
-----------+-------------------------------------------------------------
         0 |         1             9            1438       .0062587
         1 |         1            17            1214       .0140033
-------------------------------------------------------------------------
```

5.8- Regressão de Cox (análise de sobrevida)

Exemplo em Rabe-Hesketh et Evetitt, 2004

Abrir o banco c:\cursosta\heroin

Pacientes com dependência a heroína, internados em uma clínica de tratamento com metadona. O evento de interesse é abandono do tratamento. Os pacientes ainda internados no término do estudo estão registrados na variável **status** (1 se o paciente abandonou o tratamento, 0 caso contrário). As variáveis explanatórias para a saída do tratamento são dose máxima de metadona, detenção prisional e clínica onde foi internado. Estes dados foram coletados e analisados por Caplehorn e Bell (1991). Variáveis estudadas:

id: identificação do paciente

clinic: clínica de internação (1, 2)

status: variável de censura (1 - abandono, 0 - em tratamento)

time: tempo de tratamento, em dias

prison: tem registro de encarceramento (1) ou não (0)

dose: log(dose máxima de metadona)

4.4.8.1 - Apresentação dos dados

Declarando os dados como sendo na forma "st" (survival time)

- **stset time, failure(status)**

Ou

Statistics → Survival analysis ⟶ Setups & utilities ⟶ Declare data to be survival-time data

```
     failure event:  status ~= 0 & status ~= .
obs. time interval:  (0, time]
 exit on or before:  failure

--------------------------------------------------------------------------
      238  total obs.
        0  exclusions
--------------------------------------------------------------------------
      238  obs. remaining, representing
      150  failures in single record/single failure data
    95812  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =      1076
```

Resumindo os dados

- **stsum**

Ou

Statistics → Survival analysis ⟶ Setups & utilities ⟶ Declare data to be survival-time data

```
failure _d:  status
   analysis time _t:  time

         |                incidence       no. of    |------ Survival time -----|
         | time at risk      rate        subjects        25%       50%       75%
---------+-------------------------------------------------------------------
   total |        95812   .0015656           238        212       504       821
```

São 238 pacientes, com tempo mediano de "sobrevida" de 504 dias. Se a taxa de incidência (hazard ratio) for constante, é estimada como 0,0016 abandonos por dia, que corresponde a 150 abandonos/95812 dias.

Pode-se realizar a análise para cada clínica:

- **stsum, by(clinic)**

```
        failure _d:  status
  analysis time _t:  time

        |                incidence       no. of    |------ Survival time -----|
clinic  | time at risk      rate        subjects        25%       50%       75%
--------+-------------------------------------------------------------------------
      1 |        59558   .0020484           163        192       428       652
      2 |        36254   .0007723            75        280         .         .
--------+-------------------------------------------------------------------------
  total |        95812   .0015656           238        212       504       821
```

4.4.8.2- Curvas Kaplan-Meier

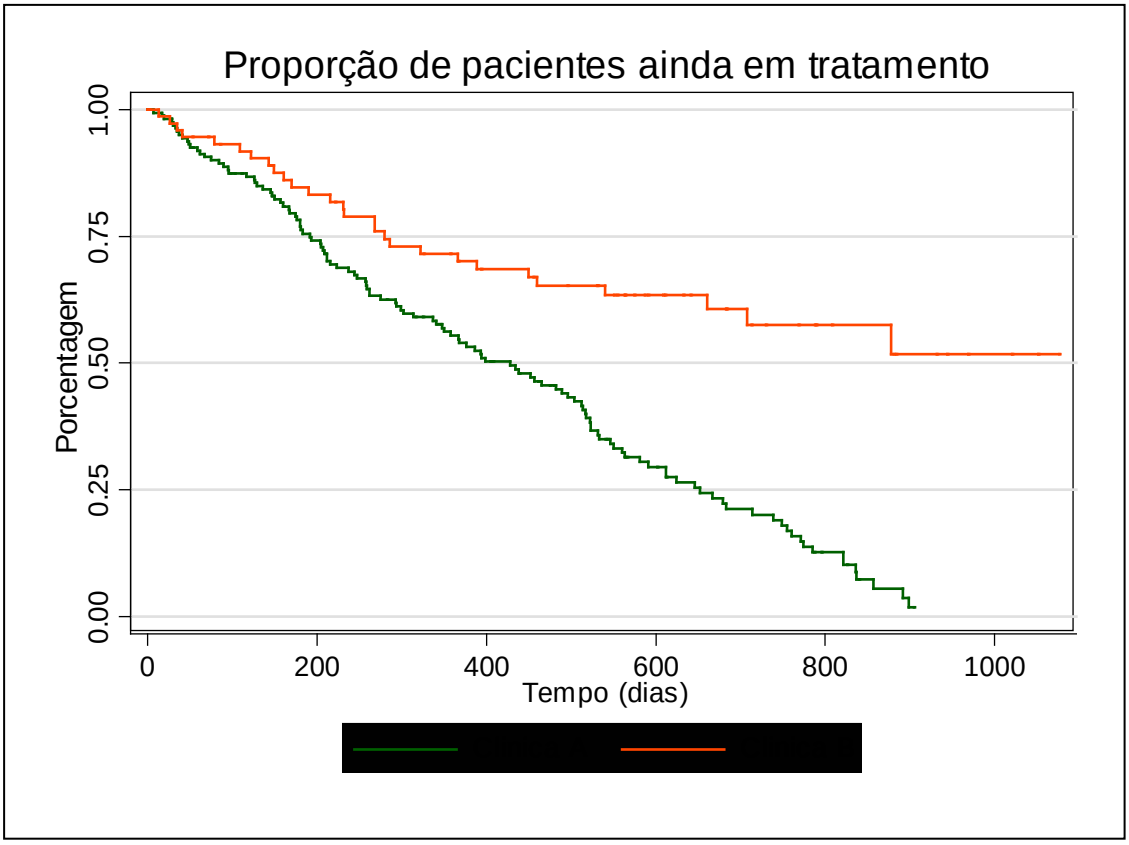
Construindo gráficos das curvas Kaplan-Meier

- **sts graph, by(clinic) ytitle(Porcentagem) yscale(range(0 1) outergap(.25)) xtitle(Tempo(dias)) title(Proporção de pacientes ainda de tratamento) legend(order(1 "Clinica A" 2 "Clinica B"))**

Ou

Statistics → Survival analysis → Summary statistics, tests & tables →

Graph survival & cumulative hazards function



Realizando o teste para igualdade das funções de sobrevida:

- **sts test clinic**

Ou

Statistics → Survival analysis ⟶ Summary statistics, tests & tables ⟶

Test equality of survivor functions

```
         failure _d:  status
   analysis time _t:  time


Log-rank test for equality of survivor functions (teste Mantel-Cox)
------------------------------------------------

       |  Events
clinic |  observed       expected
-------+-------------------------
1      |       122          90.91
2      |        28          59.09
-------+-------------------------
Total  |       150         150.00

             chi2(1) =      27.89
             Pr>chi2 =     0.000
```

5.8.3 – Regressão de Cox

- **stcox clinic**

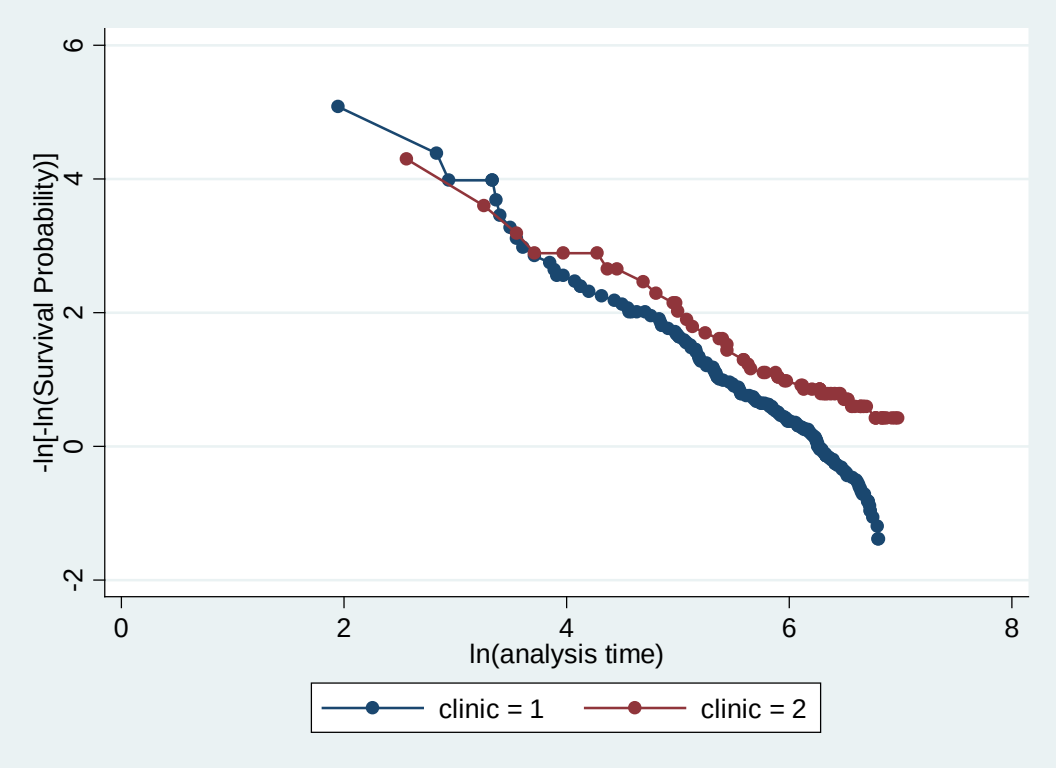
Ou

Statistics → Survival analysis → Regression models → **Cox proportional hazards model**

```
failure _d:  status
   analysis time _t:  time

Iteration 0:   log likelihood =  -705.6619
Iteration 1:   log likelihood = -690.57156
Iteration 2:   log likelihood = -690.20742
Iteration 3:   log likelihood = -690.20658
Refining estimates:
Iteration 0:   log likelihood = -690.20658

Cox regression -- Breslow method for ties

No. of subjects =          238                     Number of obs   =       238
No. of failures =          150
Time at risk    =        95812
                                                   LR chi2(1)      =     30.91
Log likelihood  =   -690.20658                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t |
          _d | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      clinic |   .3416238   .0726424    -5.05   0.000     .2251904    .5182585
------------------------------------------------------------------------------
```

Uma questão importante é se a proporcionalidade constante das *hazards* não é violada quando da comparação entre as clínicas ou da comparação entre prisioneiros e não prisioneiros. As *hazards ratios* devem ser constantes no tempo.

- **stphplot,by(clinic)**



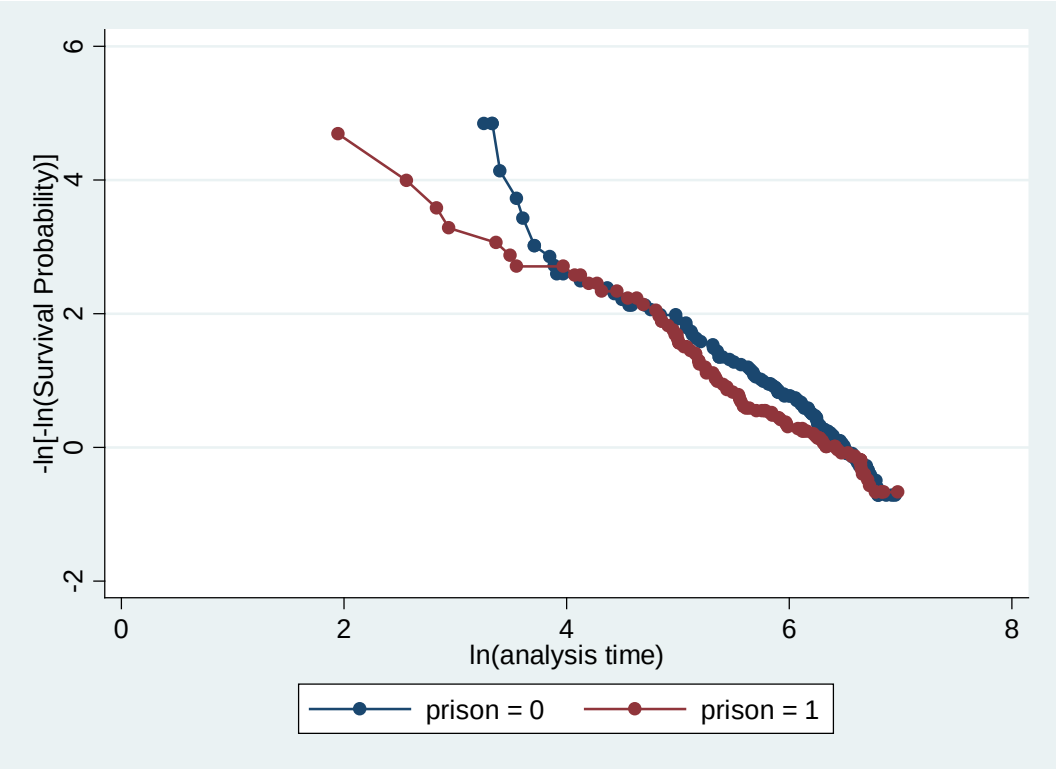
Utilizando clinicas como estrato e as outras variáveis como explanatórias

- **stcox dose prison, strata(clinic)**

```
. stcox dose prison,strata(clinic)

         failure _d:  status
   analysis time _t:  time

Iteration 0:   log likelihood = -614.68365
Iteration 1:   log likelihood = -597.73516
Iteration 2:   log likelihood =   -597.714
Refining estimates:
Iteration 0:   log likelihood =   -597.714

Stratified Cox regr. -- Breslow method for ties

No. of subjects =          238                     Number of obs   =       238
No. of failures =          150
Time at risk    =        95812
                                                   LR chi2(2)      =     33.94
Log likelihood  =     -597.714                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
      _t |
      _d | Haz. Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    dose |   .9654655   .0062418     -5.436   0.000        .953309     .977777
 prison  |   1.475192   .2491827      2.302   0.021       1.059418    2.054138
------------------------------------------------------------------------------
                                                   Stratified by clinic
```

Pacientes com história de prisão tendem a abandonar o tratamento mais rapidamente do que aqueles sem história de prisão. Para cada aumento de uma unidade (1 mg) na dose de metadona, o *hazard* é multiplicado por 0,965, ou seja, maior dose de metadona implica maior tempo no tratamento. Pacientes da clínica ficam mais tempo em tratamento.

- **stphplot,by(prison)**



5.9- Regressão linear mista (multinível); xtmixed, gllamm

Exemplo em Rabe-Hesketh et Skrondal, 2008

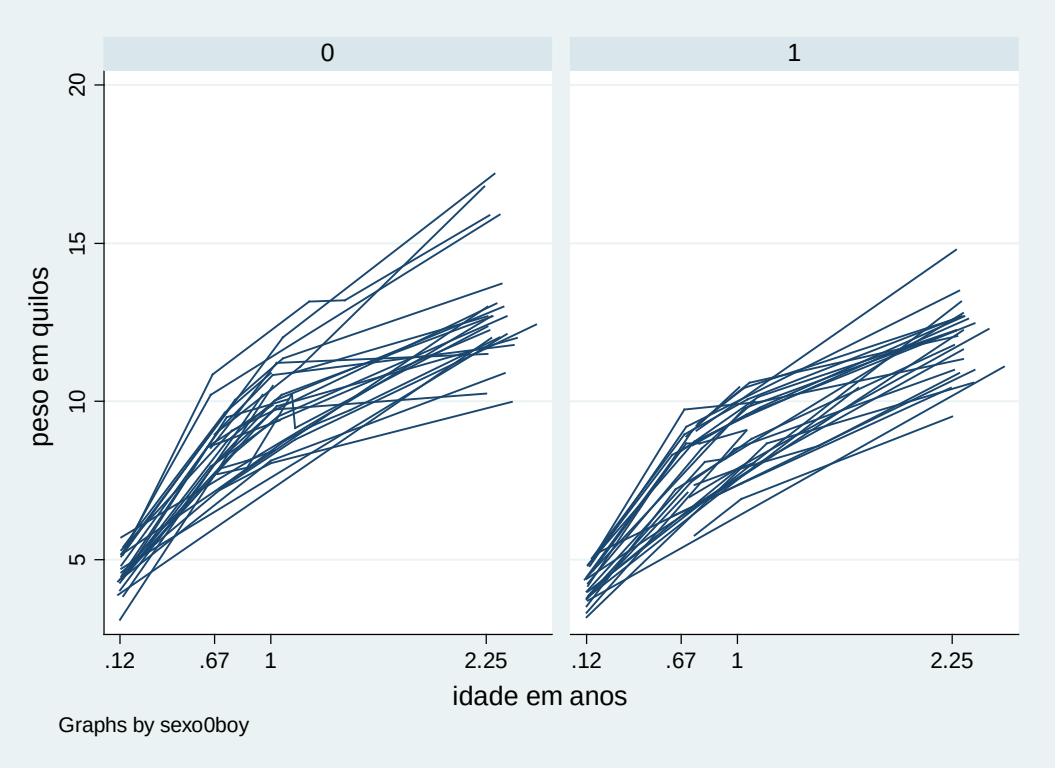
Crianças asiáticas de uma comunidade britânica, pesadas em quatro ocasiões, aproximadamente nas idades de 4 semanas (0,12 anos) e depois nas idades de 8, 12 e 27 meses (0,67, 1 e 2,25 anos). Corresponde a uma amostra casual simples de 12% de um banco maior, asian.dat, encontrável no site do Centre for Multilevel Modelling. Statamixed.dta pode ser baixado usando

- **use http://www.stata-press.com/data/mlmus2/asian**

O objetivo é estudar as trajetórias de crescimento dos pesos à medida que as crianças aumentam de idade.

- **use "C:\cursostata \Statamixed.dta"**
- **gen idadeanos2=round(idadeanos,.01)**
- **gen idadeanos2qd= idadeanos2^2**
- **sort id idadeanos2**

twoway line pesokg idadeanos2,con(ascending)||,by(sexo0boy) xtit(idade em anos) ytit(peso em quilos) xlab(.12 .67 1 2.25)



- **tab occ**

```
. tab occ
```

occ	Freq.	Percent	Cum.
1	68	34.34	34.34
2	64	32.32	66.67
3	45	22.73	89.39
4	18	9.09	98.48
5	3	1.52	100.00
Total	198	100.00	

- **xtdes ,i(id) t(occ)**

id: 45, 258, ..., 4975 n = 68
occ: 1, 2, ..., 5 T = 5

Delta(occ) = 1 unit
Span(occ) = 5 periods
(id*occ uniquely identifies each observation)

Distribution of T_i:	min	5%	25%	50%	75%	95%	max
	1	1	2	3	4	4	5

Freq.	Percent	Cum.	Pattern
27	39.71	39.71	111..
19	27.94	67.65	11...
15	22.06	89.71	1111.
4	5.88	95.59	1....
3	4.41	100	11111
68	100		XXXXX

- **gen idadeanos2qd=idadeanos2^2**

Modelo completo, sem interação:

$$y_{ij} = \beta_1 + \beta_2 x_{ij} + \beta_3 x_{ij}^2 + \beta_4 w_j + \varsigma_{1j} + \varsigma_{2j} x_{ij} + \varepsilon_{ij}$$

- **xtmixed pesokg idadeanos2 idadeanos2qd sexo0boy || id: idadeanos2, cov(unstructured) ml**

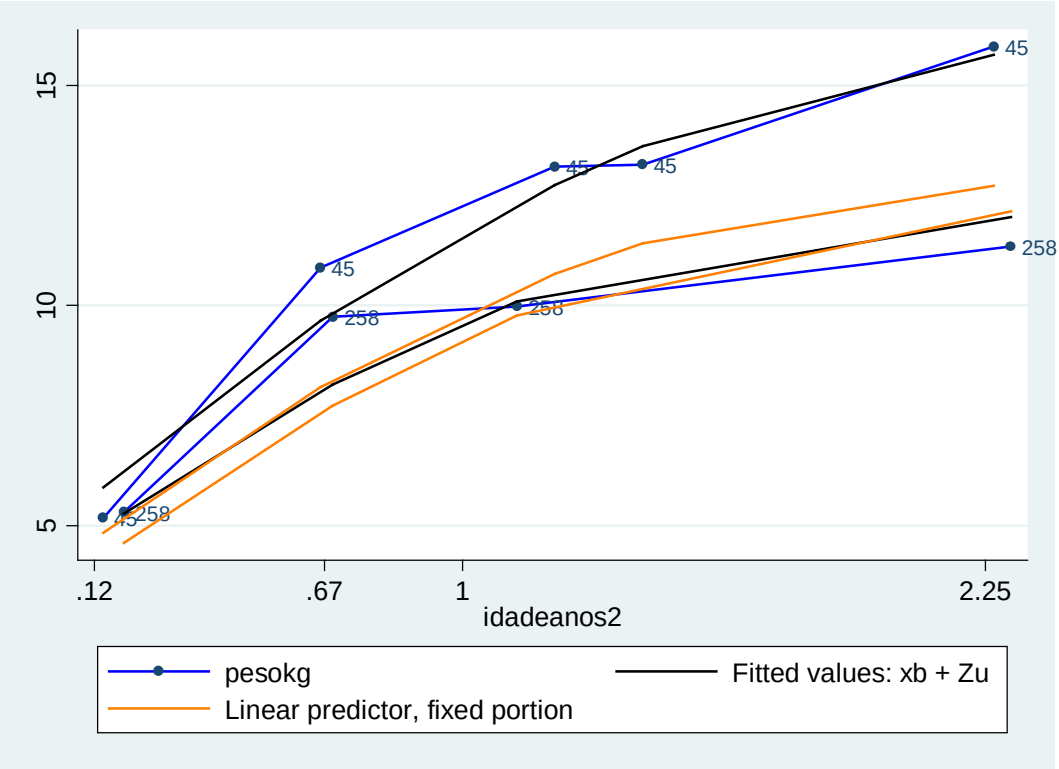
```
Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood = -253.64582
Iteration 1:   log likelihood = -253.60234
Iteration 2:   log likelihood = -253.60224
Iteration 3:   log likelihood = -253.60224

Computing standard errors:

Mixed-effects ML regression                     Number of obs      =       198
Group variable: id                              Number of groups   =        68

                                                Obs per group: min =         1
                                                               avg =       2.9
                                                               max =         5


                                                Wald chi2(3)       =   1977.39
Log likelihood = -253.60224                     Prob > chi2        =    0.0000
```

```
------------------------------------------------------------------------------
      pesokg |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  idadeanos2 |   7.695848   .2375157    32.40   0.000     7.230326     8.16137
idadeanos2qd |  -1.656517    .087744   -18.88   0.000    -1.828492   -1.484542
    sexo0boy |  -.5966719   .1962346    -3.04   0.002    -.9812845   -.2120592
       _cons |   3.793285   .1653125    22.95   0.000     3.469279    4.117291
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Unstructured             |
                sd(idadea~2) |   .5110846   .0870337      .3660495    .7135852
                   sd(_cons) |    .595347   .1286601      .3897781    .9093329
         corr(idadea~2,_cons) |   .1535116   .3214146     -.4545577    .6639807
-----------------------------+------------------------------------------------
                sd(Residual) |   .5705276   .0495085       .481296    .6763026
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(3) =   104.53   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.

. estat recovariance

Random-effects covariance matrix for level id

             | idadean~2      _cons
-------------+----------------------
  idadeanos2 |  .2612075
       _cons |  .0467094    .354438
```

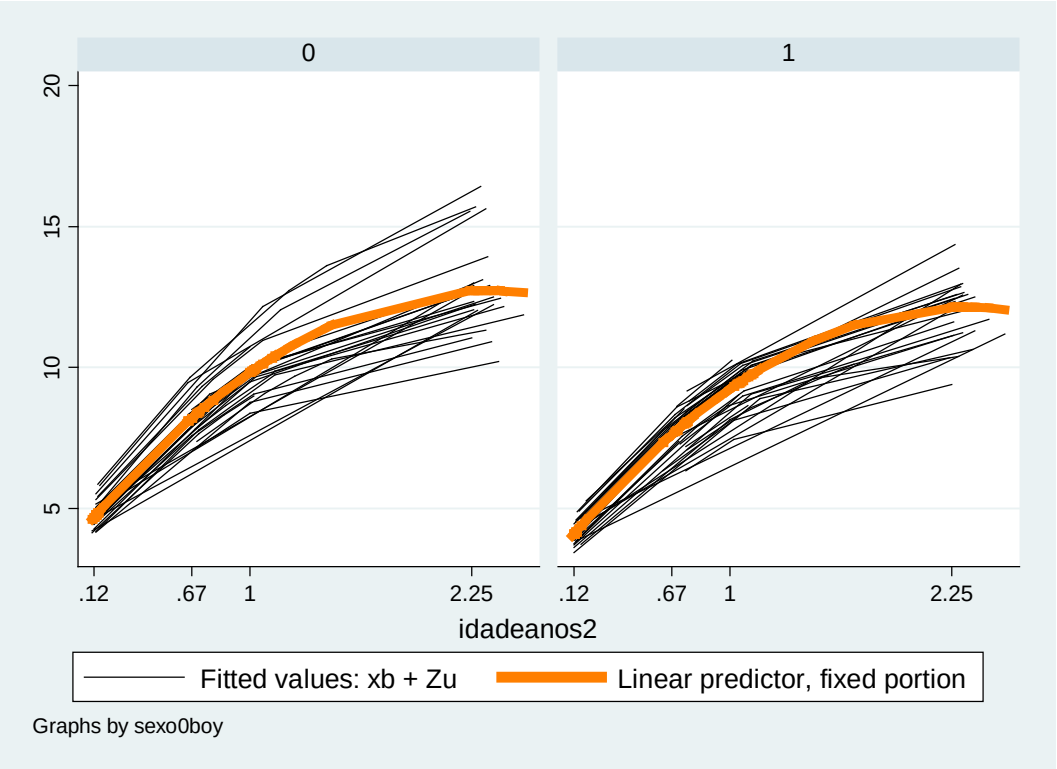
- **predict xb,xb**
- **predict stdp,stdp**
- **predict fitted,fitted**
- **predict residuals,residuals**
- **predict rstandard,rstandard**
- **predict reffects*,reffects**
- **gen xblower=xb-1.96* stdp**
- **gen xbupper=xb+1.96* stdp**

- **findit linkplot**

- **linkplot pesokg fitted xb idadeanos2 if id==45 |id==258,link(id) lcolor(blue black orange) mlabel(id) msymbol(o none none) xlab(.12 .67 1 2.25)**



- **twoway line fitted idadeanos2 ,sort(id idadeanos2) lcolor(black) lwidth(thin) connect(ascending)|| line xb idadeanos2,sort(idadeanos2)lwidth(verythick) lcolor(orange)||,xlab(.12 .67 1 2.25) by(sexo0boy)**



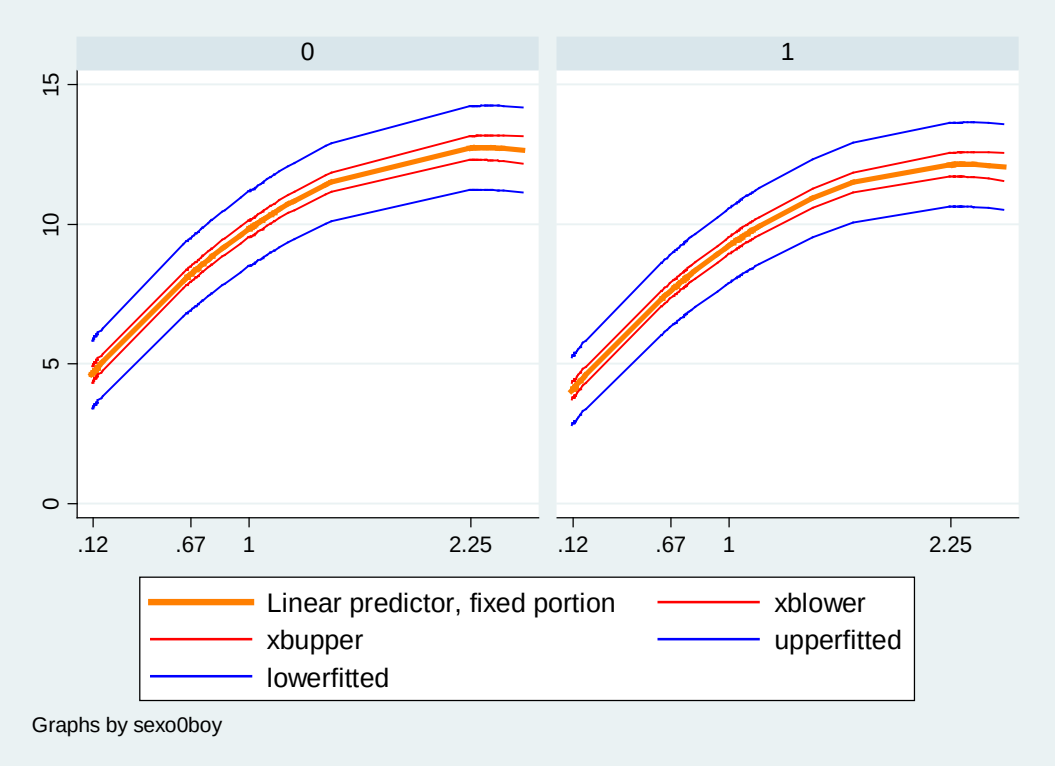
var(a+bx)= var(a)+2*cov(a,b)*x+var(b)*(x^2)

erro padrão(a+bx)= raiz quadrada [var(a+bx)]

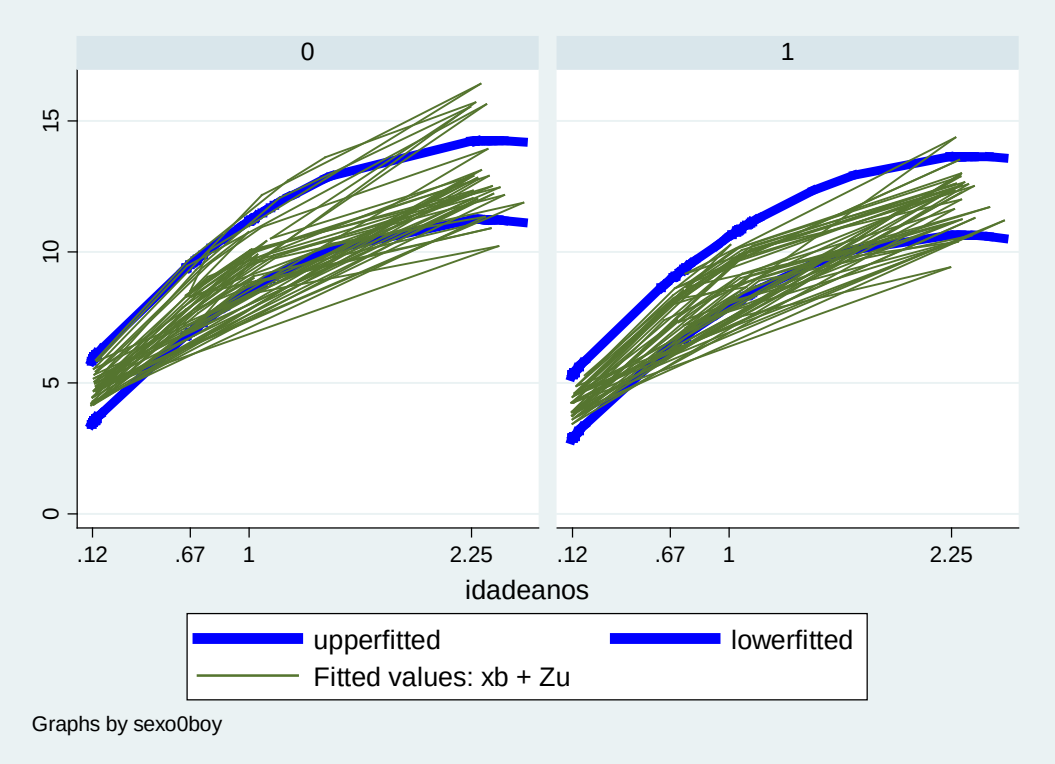
- **gen upperfitted= xb+1.96*sqrt(.354438+2* .0467094*idadeanos2+.2612075*.2612075^2)**

- **gen lowerfitted= xb-1.96*sqrt(.354438+2* .0467094*idadeanos2+.2612075*.2612075^2)**

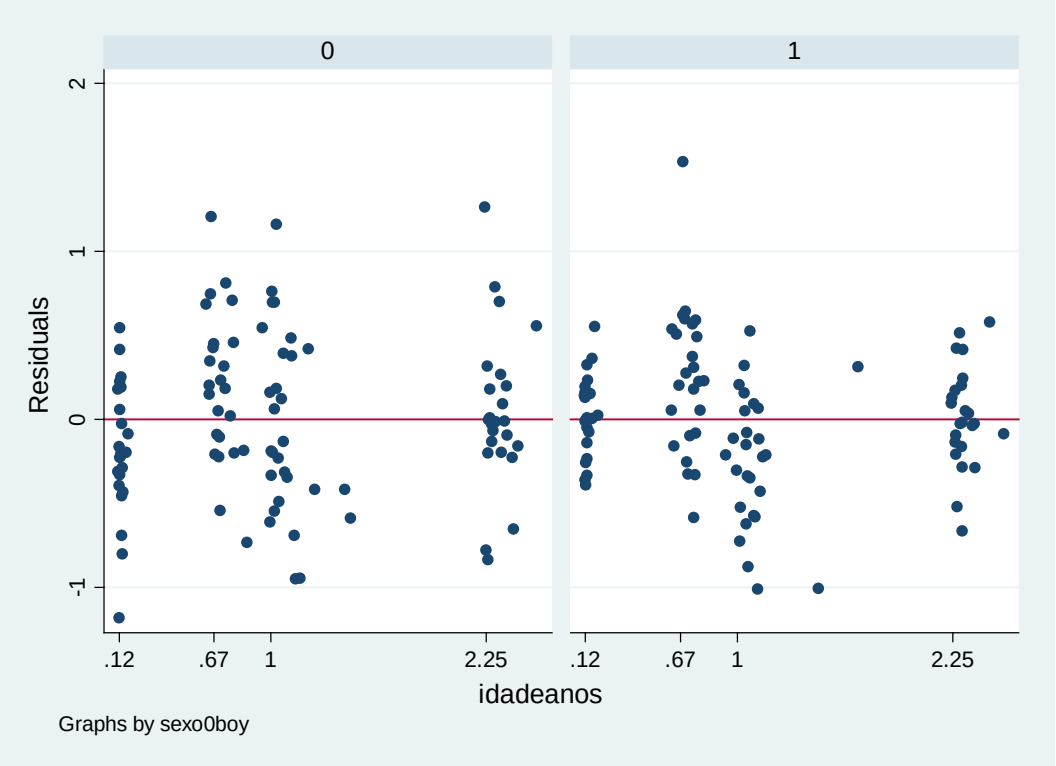
- **twoway line xb idadeanos2, sort(idadeanos2) lwidth(thick) lcolor(orange)||line xblower idadeanos,sort(idadeanos) lcolor(red)|| line xbupper idadeanos, sort(idadeanos) lcolor(red)||line upperfitted idadeanos, sort(idadeanos) lcolor(blue)||line lowerfitted idadeanos, sort(idadeanos) lcolor(blue)||,xlab(.12 .67 1 2.25) by(sexo0boy)**



- **twoway line upperfitted idadeanos,sort(idadeanos)lwidth(vthick)lcolor(blue)||line lowerfitted idadeanos, lwidth(vthick)sort(idadeanos) lcolor(blue)||line fitted idadeanos||, xlab(.12 .67 1 2.25) by(sexo0boy)**

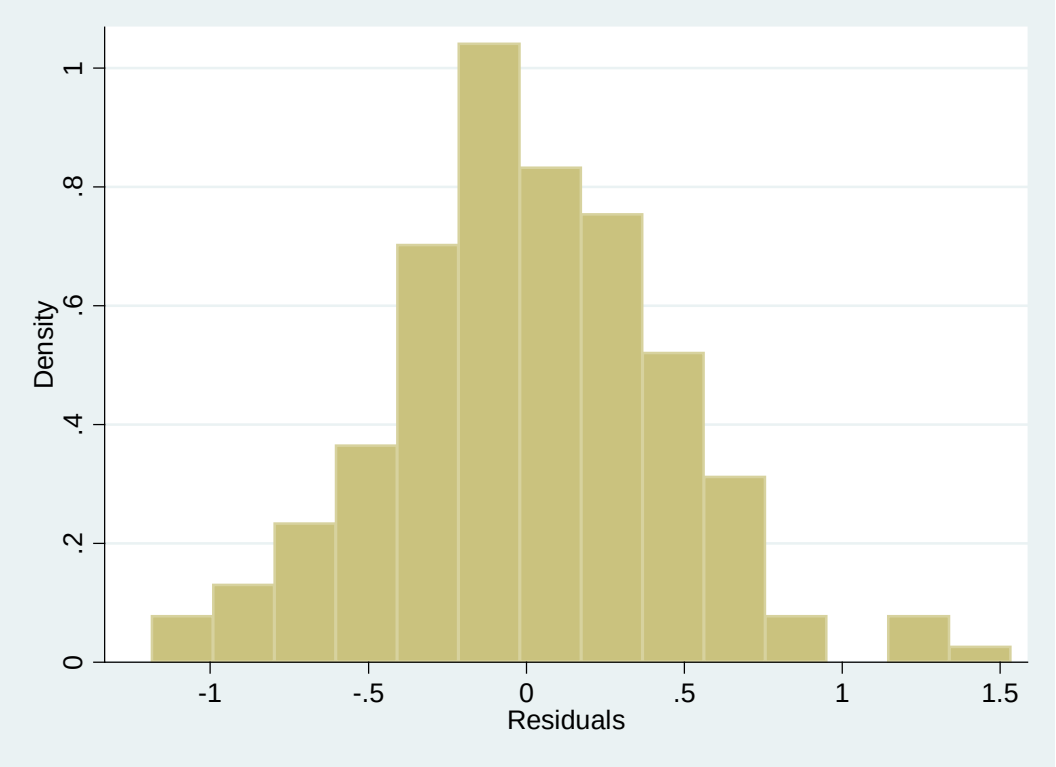


- **scatter residuals idadeanos,yline(0) by(sexo0boy) xlab(.12 .67 1 2.25))**

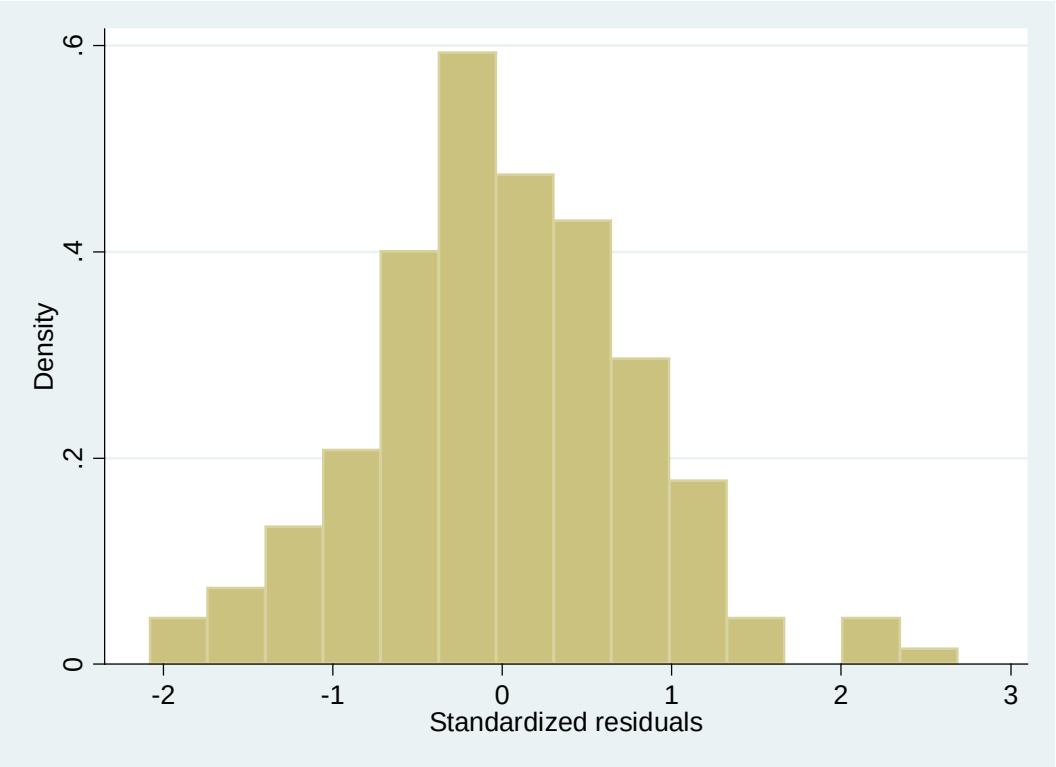


- **histogram residuals**

(bin=14, start=-1.1849141, width=.19417079)

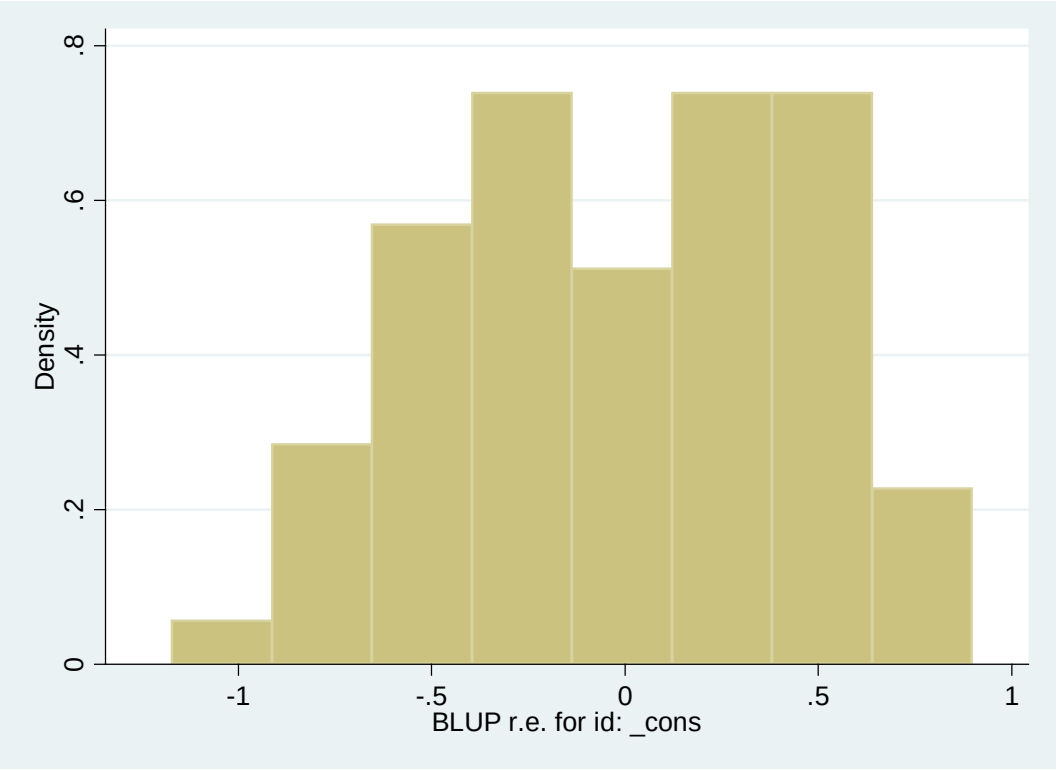


- **histogram rstandard (bin=14, start=-2.0768745, width=.34033551)**

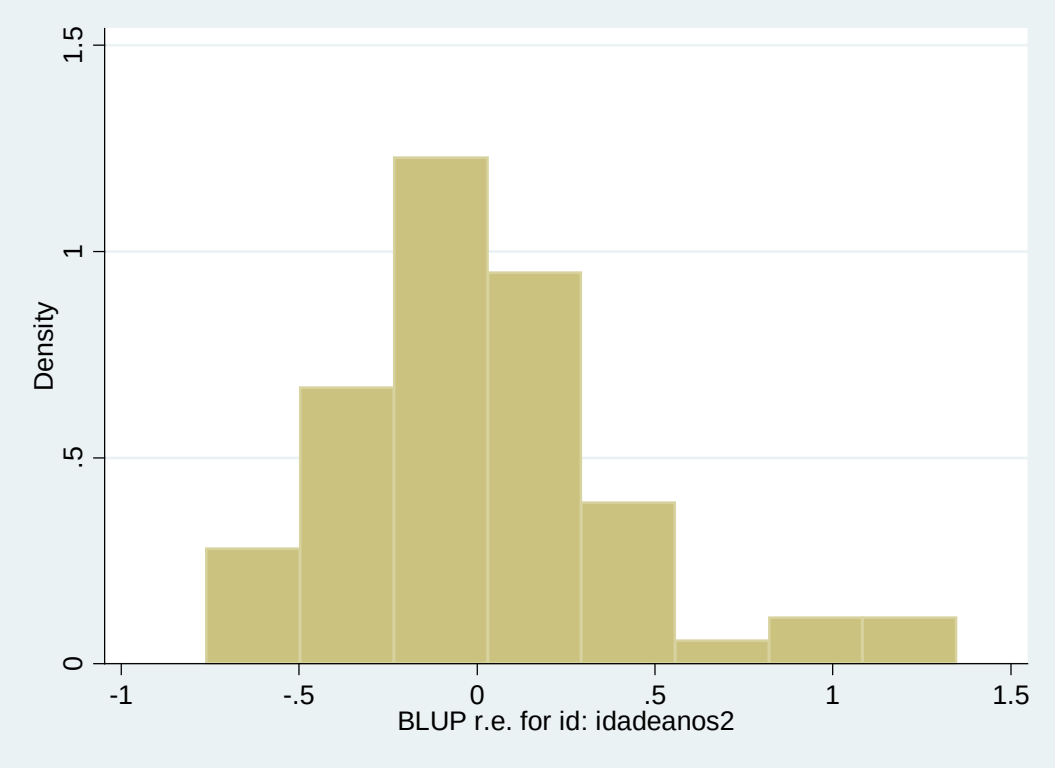


- **egen pickone=tag(id)**

- **`histogram reffects2 if  pickone==1 (bin=8,start=-1.1705161, width=.25864983)`**



- **histogram reffects1 if pickone==1 (bin=8, start=-.76046056, width=.2635813)**



5.10- Regressão logística mista (multinível); xtmelogit, gllamm

- **ssc install gllamm,replace**
- **ssc install ci_marg_mu, replace (Stata 10 ou mais)**

Material em Rabe-Hesketh et Skrondal, 2008, usando o banco HTTP://www.stata-press.com/data/mlmus2/toenail

- **use c:/cursostata/unha.dta**

Ensaio clínico com 387 pacientes com infecção na unha do dedo maior do pé. Foram testados dois medicamentos orais, itraconazole= 0 eterbinafine= 1, com sete visitas médicas, semanas 0, 4, 8, 12, 24, 36 e 48. A resposta foi a condição da infecção, nenhuma ou fraca= 0 moderada ou grave= 1.

- **xtdescribe,i(patient) t(semana) patterns(1000)**

```
patient:  1, 2, ..., 383        n =           294
semana:  0, 4, ..., 48 T =            7
Delta(semana) = 1 unit
Span(semana)  = 49 periods
(patient*semana uniquely identifies     each observation)

Distribution of T_i:   min    5%   25%    50%  75%    95%   max
                         1     3     7      7      7      7     7

Freq. Percent Cum.   Pattern*
----------------------+---------------
224 76.19   76.19   1111..1..1..1
21    7.14   83.33   1111..1.....1
10    3.40   86.73   1111.....1..1
6     2.04   88.78   111.........
```

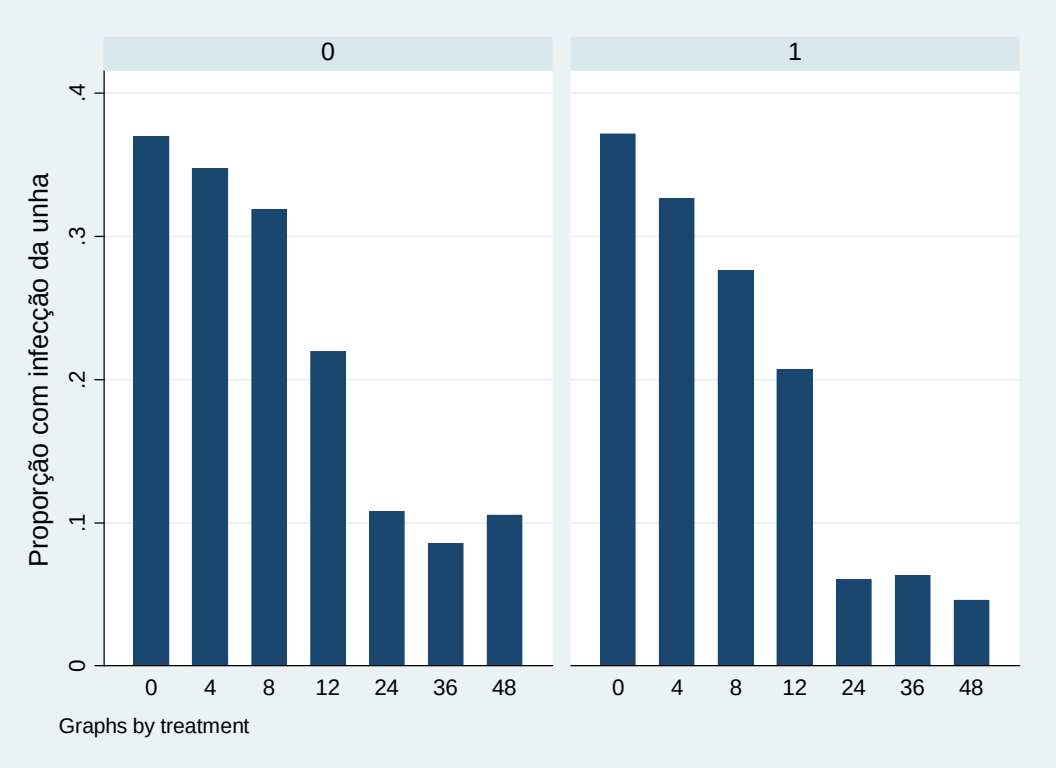
5 1.70 90.48 1.....
5 1.70 92.18 1111..1.....
4 1.36 93.54 1111.....
3 1.02 94.56 11.....
3 1.02 95.58 111...1..1..1
3 1.02 96.60 1111..1..1...
2 0.68 97.28 111...1.....
2 0.68 97.96 1111.....1
1 0.34 98.30 1.11..1..1..1
1 0.34 98.64 11....1.....
1 0.34 98.98 11.1..1..1...
1 0.34 99.32 11.1..1..1..1
1 0.34 99.66 111.....1..1
1 0.34 100 111...1.....1

----------+----------

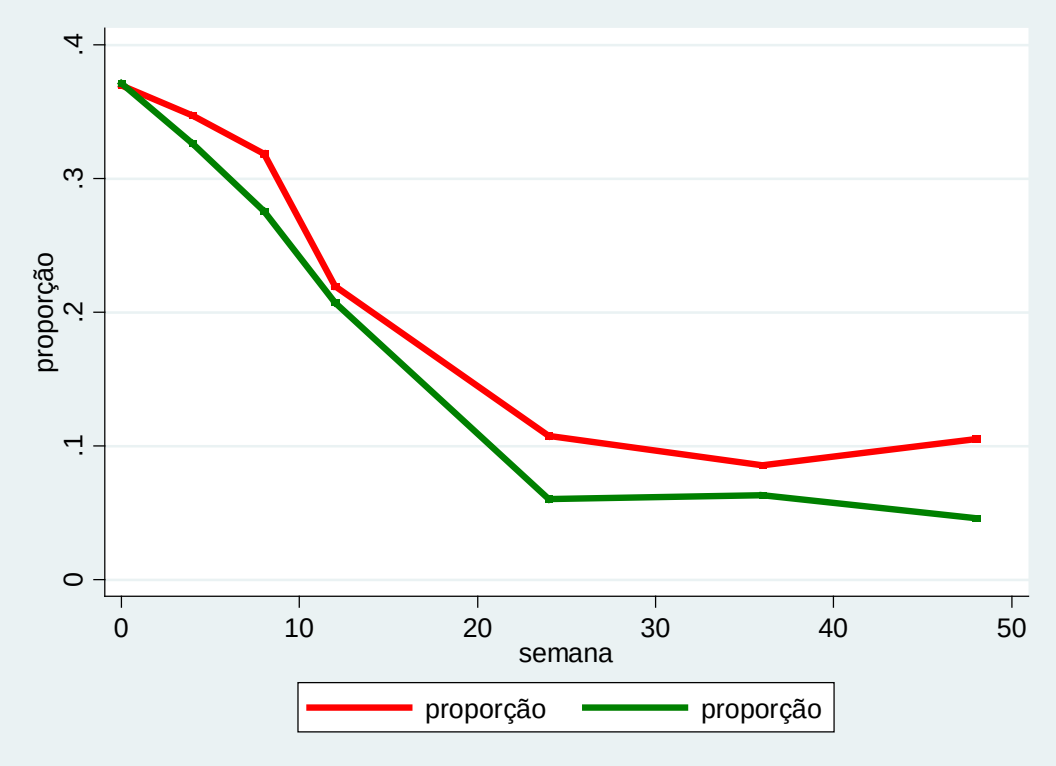
294 100.00 XXXX..X..X..X

***Each column represents 4 period**

- **graph bar (mean) outcome,over(semana) by(treatment) ytitle(Proporção com infecção da unha)**



line proporção semana if treatment==0,sort(semana)lcol-or(red)lwidth(thick)||line proporção semana if treatment==1,sort(semana)lcol-or(green)lwidth(thick)



- **egen proporção=mean(outcome),by(treatment semana)**
- **gen treatmentXsemana= treatment*semana**
- **gllamm outcome treatment semana treatmentX-semana,i(patient)link(logit)fam(binomial) nip(30)adapt**

```
Running adaptive quadrature
Iteration 0:    log likelihood = -835.21915
Iteration 1:    log likelihood = -697.01263
Iteration 2:    log likelihood = -661.97918
Iteration 3:    log likelihood = -634.56169
Iteration 4:    log likelihood = -624.82023
Iteration 5:    log likelihood = -623.92864
Iteration 6:    log likelihood = -623.89596
Iteration 7:    log likelihood = -623.89581


Adaptive quadrature has converged, running Newton-Raphson
Iteration 0:   log likelihood = -623.89581
Iteration 1:   log likelihood = -623.89557
Iteration 2:   log likelihood = -623.89554
```

```
number of level 1 units = 1908
number of level 2 units = 294

Condition Number = 87.682849

gllamm model

log likelihood = -623.89554
```

outcome	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
treatment	-.1148057	.581708	-0.20	0.844	-1.254932	1.025321
semana	-.1011043	.0114984	-8.79	0.000	-.1236408	-.0785677
treatmentX~a	-.0403507	.0179702	-2.25	0.025	-.0755716	-.0051298
_cons	-1.632857	.4335199	-3.77	0.000	-2.48254	-.7831734

```
 Variances and covariances of random effects

***level 2 (patient)

var(1): 16.162764 (3.0829554)
```

- **gllamm ,eform**

```
number of level 1 units = 1908
number of level 2 units = 294

Condition Number = 87.682849
gllamm model
log likelihood = -623.89554
```

outcome	exp(b)	Std. Err.	z	P>z	[95% Conf.	Interval]
treatment	.8915394	.5186156	-0.20	0.844	.2850951	2.78799
semana	.9038388	.0103927	-8.79	0.000	.8836972	.9244394
treatmentX~a	.9604526	.0172595	-2.25	0.025	.9272133	.9948834

```
Variances and covariances of random effects

***level 2 (patient)

var(1): 16.162764 (3.0829554)
```

- **gllapred marginalprob, mu marginal**

```
(mu will be stored in marginalprob)
```

- **gllapred cmuindividual, mu**

```
(mu will be stored in cmuindividual)
Non-adaptive log-likelihood: -624.11432
-623.8953  -623.8955  -623.8955
log-likelihood:-623.89553
```

- **ci_marg_mu lower upper,dots**

.....

11.498 seconds = .19163333 minutes = .00319389 hours

.....

> .....

> .....

> .....

> .....

> .....

> .....

> .....

> .....

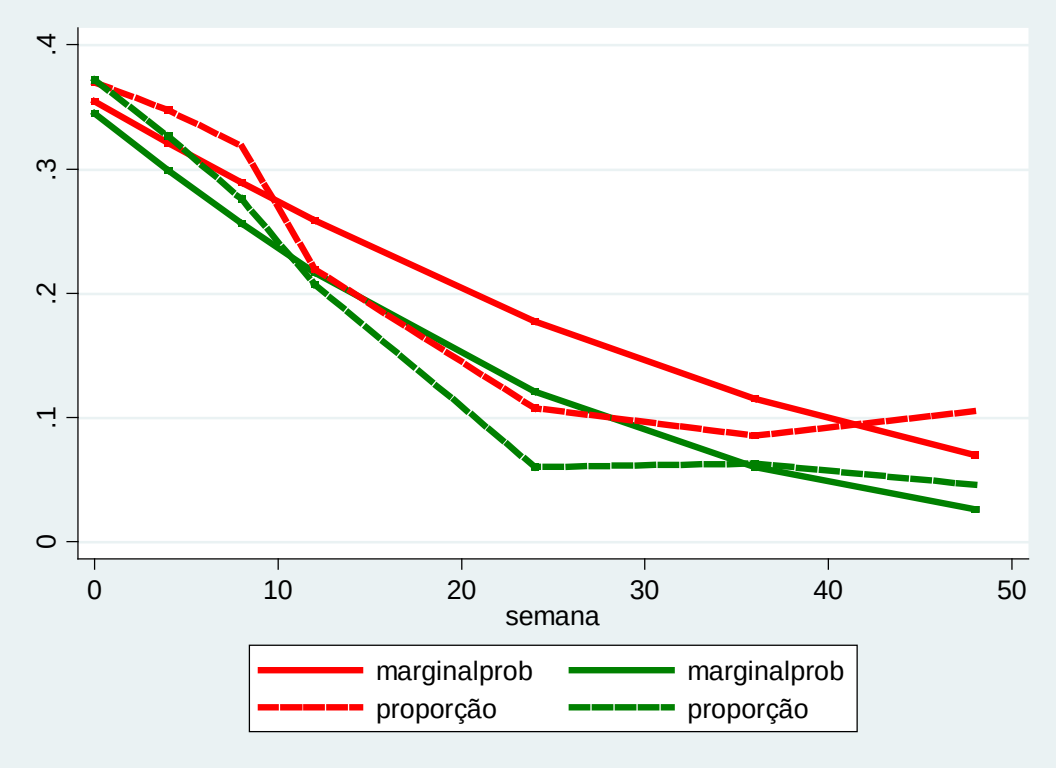
> .....

> .....

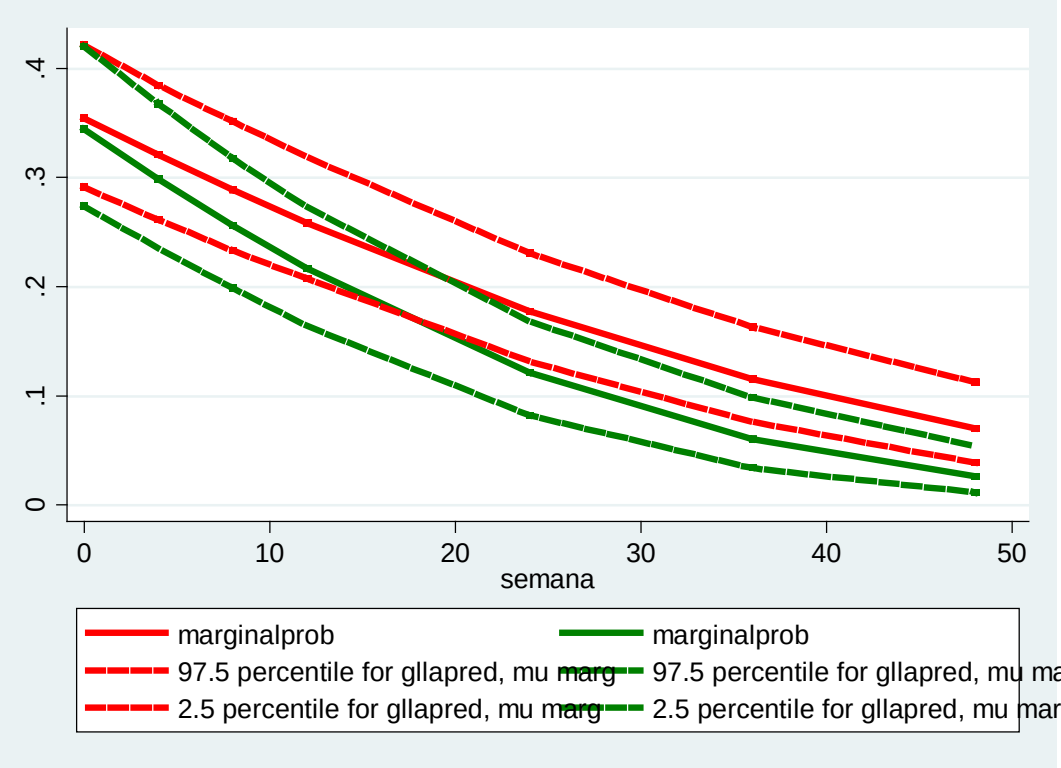
> .....

276.76 seconds = 4.6126667 minutes = .07687778 hours

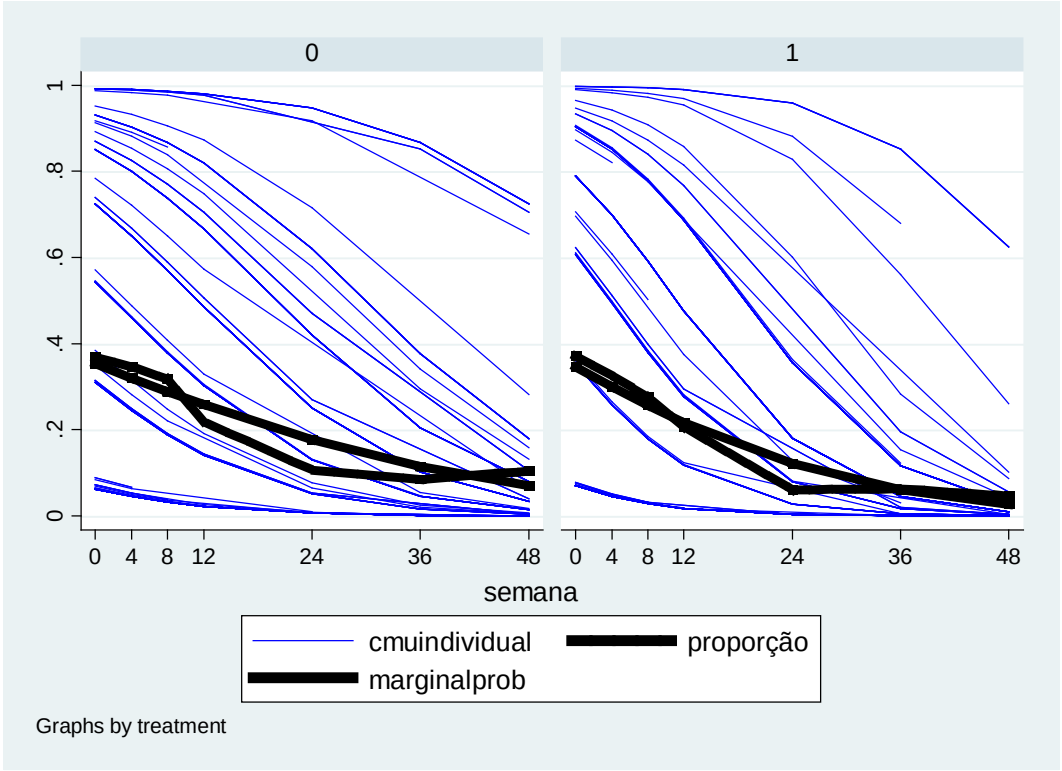
line marginalprob semana if treatment==0,sort(semana) lcolor(red)con(ascending)lwidth(thick)||line marginalprob semana if treatment==1,sort(semana) lwidth(thick)lcolor(green)con(ascending)||line proporção semana if treatment==0,sort(semana)lpat(dash)lcolor(red)lwidth(thick)||line proporção semana if treatment==1,sort(semana)lcolor(green)lwidth(thick)lpat(dash)


0
.1
.2
.3
.4
0
10
20
30
40
50
semana
marginalprob
proporção
marginalprob
proporção

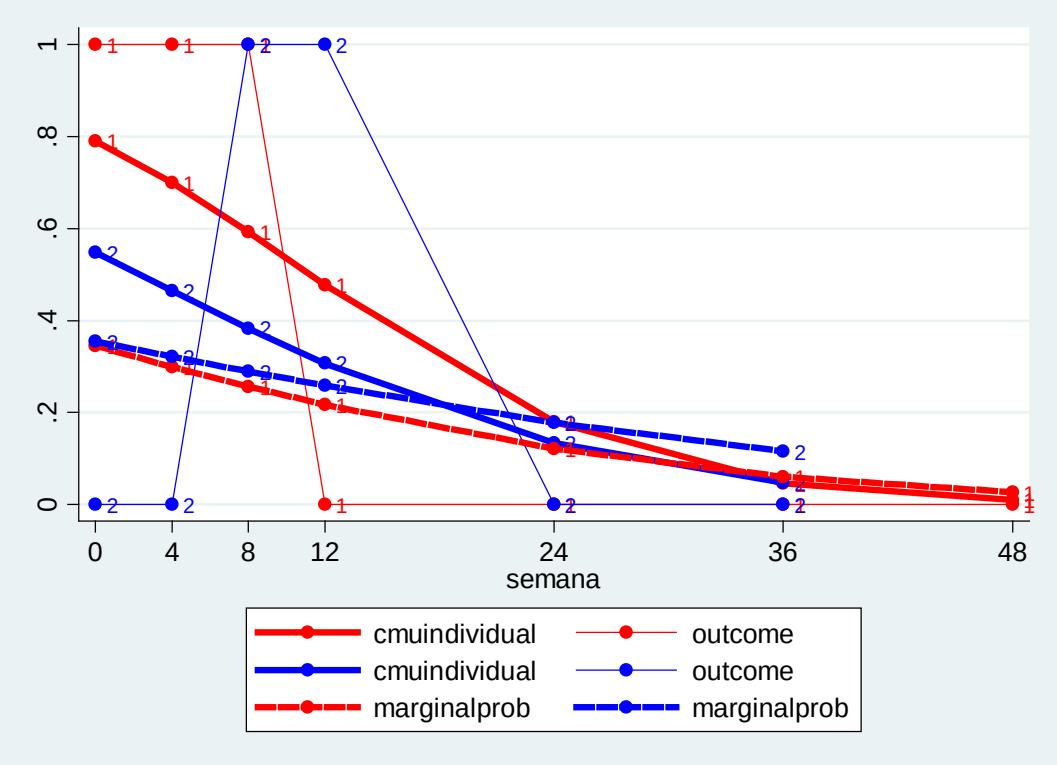
line marginalprob semana if treatment==0,sort(semana) lcolor(red)con(ascending)lwidth(thick)||line marginalprob semana if treatment==1,sort(semana) lwidth(thick)lcolor(green)con(ascending)||line upper semana if treatment==0,sort(semana)lpat(dash)lcolor(red)lwidth(thick)||line upper semana if treatment==1,sort(semana)lcolor(green)lwidth(thick)lpat(dash)||line lower semana if treatment==0,sort(semana)lpat(dash)lcolor(red)lwidth(thick)||line lower semana if treatment==1,sort(semana)lcolor(green)lwidth(thick)lpat(dash)||



linkplot cmuindividual semana ,link(patient) sort(patient semana) msym-bol(i)lcolor(blue)con(ascending)lwidth(thin)xlab(0 4 8 12 24 36 48)ylab(0 .2 .4 .6 .8 1) by(treatment) plot(line proporção marginalprob semana,sort(semana) lwidth(vthick vthick) lcolor(black black)lpat(dash solid))



scatter cmuindividual outcome semana if patient==1,sort(semana)con(l l) lwidth(thick thin)mlab(patient patient) lcolor(red red)mcolor(red red)mlabcolor(red red)||scatter cmuindividual outcome semana if patient==2 , sort(semana) lwidth(thick thin) lcolor(blue blue)mlabcolor(blue blue)mlabel(patient patient) mcolor(blue blue)con(l l)||scatter marginalprob semana if patient==1, sort(semana) con(l) lwidth(thick)mlab(patient) lcolor(red)mcolor(red)mlabcolor(red)lpat(dash)||scatter marginalprob semana if patient==2, sort(semana) con(l) lwidth(thick)mlab(patient) lcolor(blue)mcolor(blue)mlabcolor(blue)lpat(dash)||,xlab(0 4 8 12 24 36 48)ylab(0 .2 .4 .6 .8 1)



Paciente 1 pertence ao tratamento 1 (vermelho)
Paciente 2 pertence ao tratamento 0 (azul)

5.11- Regressão de Poisson mista (multinível); xtmepoisson

Exemplo do manual [XT] Stata, versão 10, usando dados de Thall & Vail, 1990: http://www.stata-press.com/data/r10/epilepsy

- **use c:cursostata/epilepsia.dta**

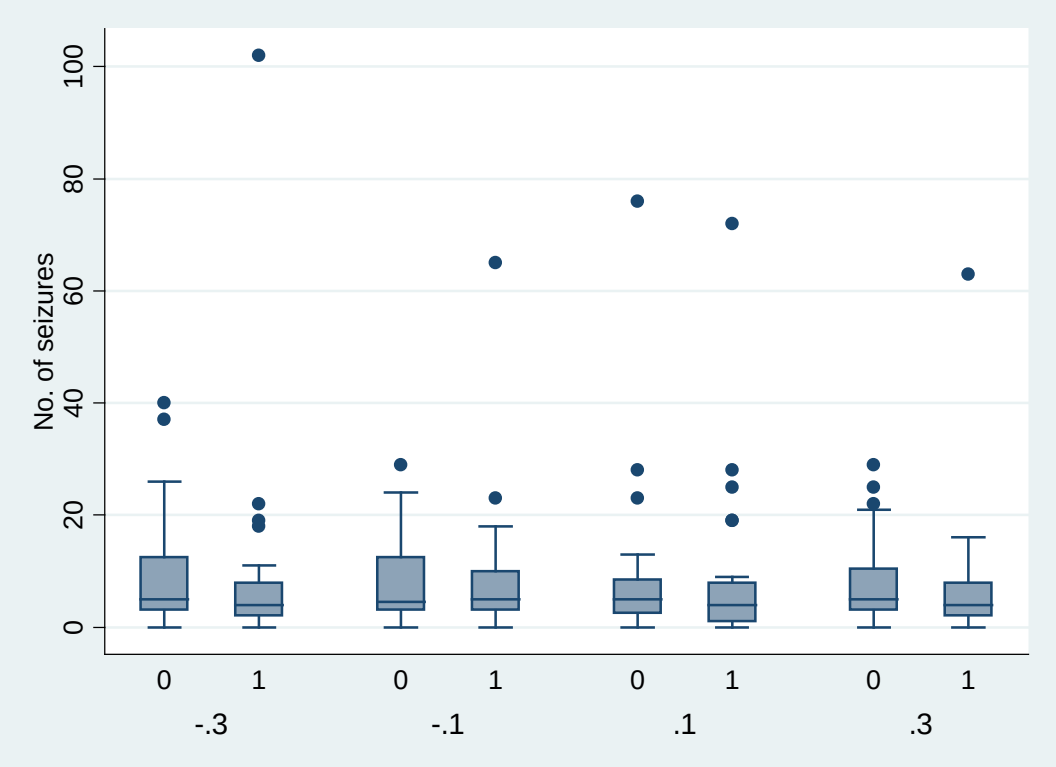
São 59 pacientes observados quanto ao número de episódios de epilepsia (seizures) ocorridos durante duas semanas antes de cada uma de quatro visitas médicas (centralizadas em -.3 -.1 .1 .3). A comparação é entre tratamento: progabide= 1 e placebo= 0

- describe

```
subject         byte    %9.0g                 Subject id: 1-59
seizures        int     %9.0g                 No. of seizures
treat           byte    %9.0g                 1: progabide; 0: placebo
visit           float   %9.0g                 Dr. visit; coded as (-.3, -.1, .1, .3)
lage            float   %9.0g                 log(age), mean-centered
lbas            float   %9.0g                 log(0.25*baseline seizures), mean-centered
lbas_trt        float   %9.0g                 lbas/treat interaction
v4              byte    %8.0g                 Fourth visit indicator
visita          float   %9.0g
treatXvisit     float   %9.0g
mu              float   %9.0g                 Predicted mean
média           float   %9.0g
```

```
. xtdescribe ,i( subject) t( visit)
time variable must contain only integer values
r(451);

. gen visita= visit

. recode visita -.3=1  -.1=2 .1=3 .3=4
(visita: 236 changes made)

. xtdescribe ,i( subject) t( visita)

 subject:  1, 2, ..., 59                                    n =         59
  visita:  1, 2, ..., 4                                     T =          4
           Delta(visita) = 1 unit
           Span(visita)  = 4 periods
           (subject*visita uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         4       4       4         4         4       4       4

     Freq.  Percent    Cum.  |  Pattern
------------------------------+---------
        59    100.00  100.00  |  1111
------------------------------+---------
        59    100.00          |  XXXX
```

- **graph box seizures,over(treat) over(visit)**



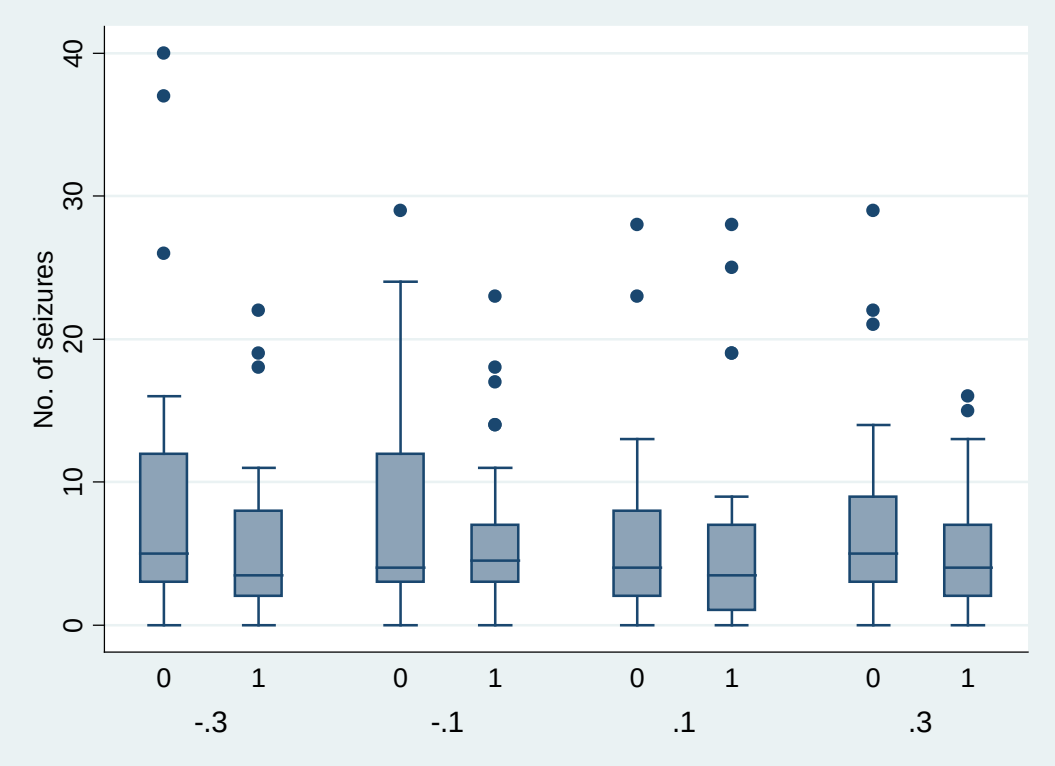
- **list subject seizures if seizures>60**

	subject	seizures
49.	49	102
108.	49	65
143.	25	76
167.	49	72
226.	49	63

list subject seizures if subject==25| subject==49

	subject	seizures
25.	25	18
49.	49	102
84.	25	24
108.	49	65
143.	25	76
167.	49	72
202.	25	25
226.	49	63

- **graph box seizures if subject!=49&subject!=25,over(treat) over(visit)**



•

- **xtmepoisson seizures treat visit if subject!=49 & subject!=25|| subject: visit,cov(unstructured) intpoints(9)**

```
Refining starting values:
Iteration 0:   log likelihood = -629.40063
Iteration 1:   log likelihood = -623.49844
Iteration 2:   log likelihood = -622.77124
Performing gradient-based optimization:
Iteration 0:   log likelihood = -622.77124
Iteration 1:   log likelihood = -622.75651
Iteration 2:   log likelihood =  -622.7565

Mixed-effects Poisson regression            Number of obs     =       228
Group variable: subject                     Number of groups  =        57

Obs per group: min =         4
avg =       4.0
max =         4

Integration points =   9                    Wald chi2(2)      =      3.69
Log likelihood =  -622.7565                 Prob > chi2       =    0.1583

------------------------------------------------------------------------------
    seizures      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------------------------------------------------------------------
       treat   -.303985   .2279996    -1.33   0.182    -.750856     .142886
       visit  -.2588333   .1821949    -1.42   0.155    -.6159288    .0982623
       _cons   1.692118   .1642509    10.30   0.000     1.370192    2.014044
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
subject: Unstructured        |
                  sd(visit)  |   .6978556   .1717472      .4308032    1.130452
                  sd(_cons)  |   .8193085   .0875546      .6644842    1.010207
             corr(visit,_cons)|  -.1080117   .2518356     -.5426118    .3722052
------------------------------------------------------------------------------

LR test vs. Poisson regression:      chi2(3) =   813.39   Prob > chi2 = 0.0000
Note: LR test is conservative and provided only for reference.
```

- **xtmepoisson,irr**

```
------------------------------------------------------------------------------
    seizures |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   .7378719   .1682345    -1.33   0.182    .4719624    1.153598
```

visit .7719517 .1406457 -1.42 0.155 .540139 1.103252

- **predict mubeta,mu**

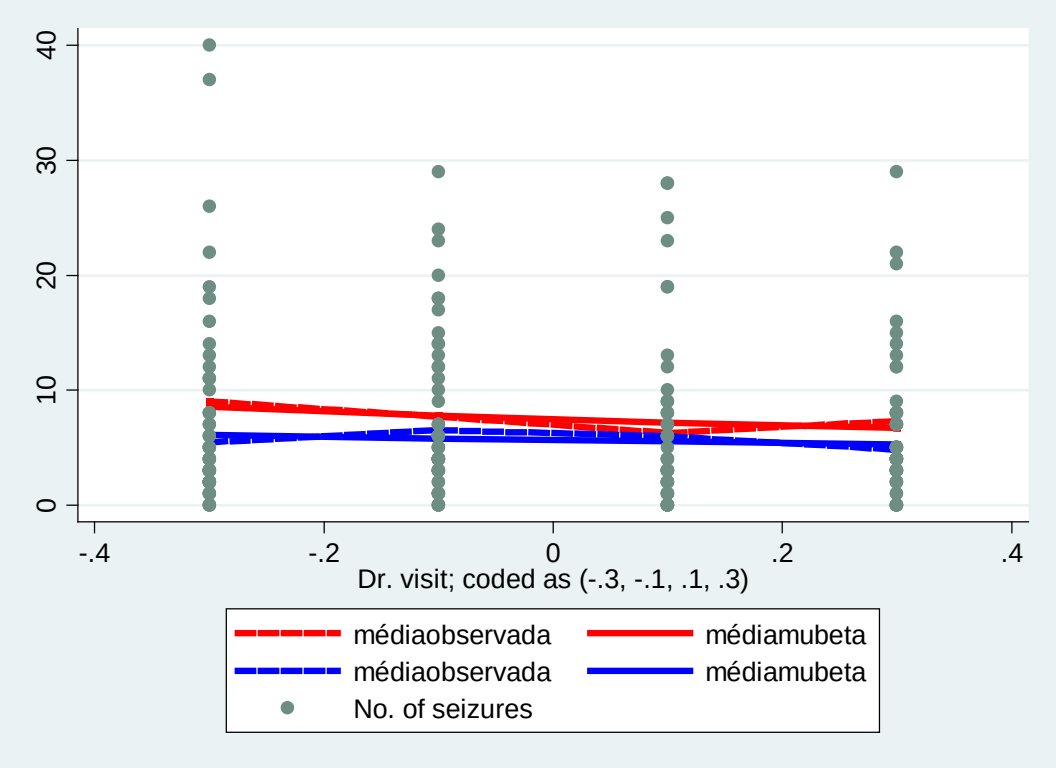
(8 missing values generated)

- **egen médiamubeta=mean(mubeta) if subject!=49&subject!=25, by(treat visit)**

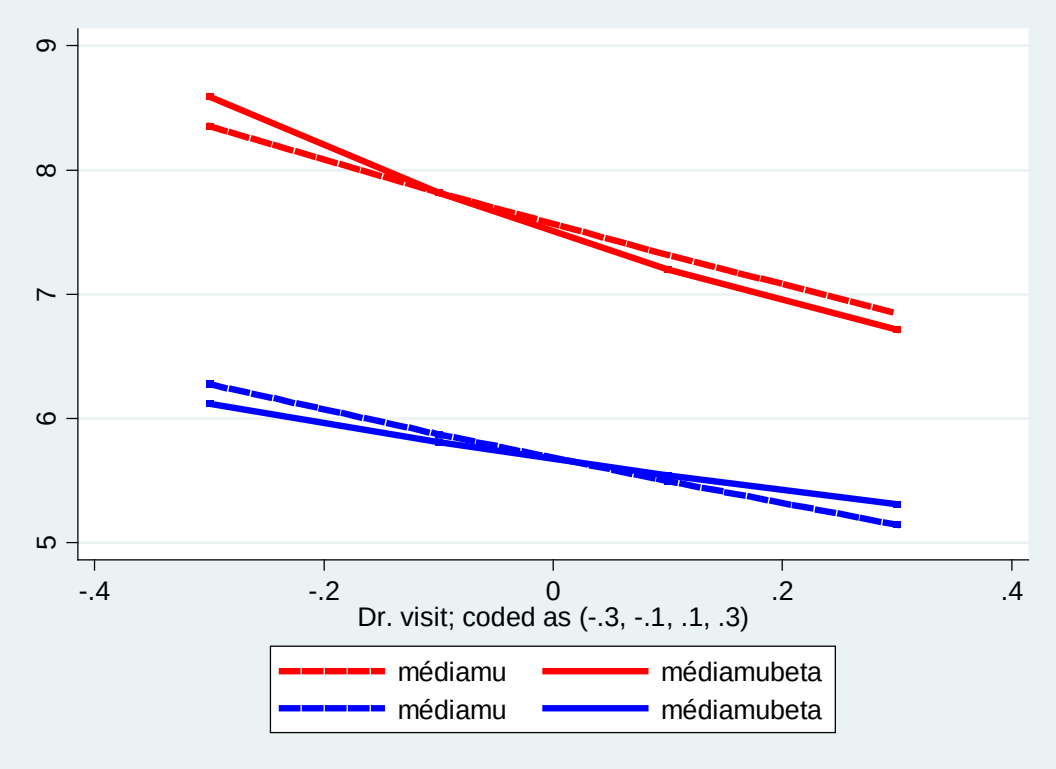
(8 missing values generated)

- **egen médiaobservada=mean(seizures) if subject!=49 &subject!=25, by(treat visit)**

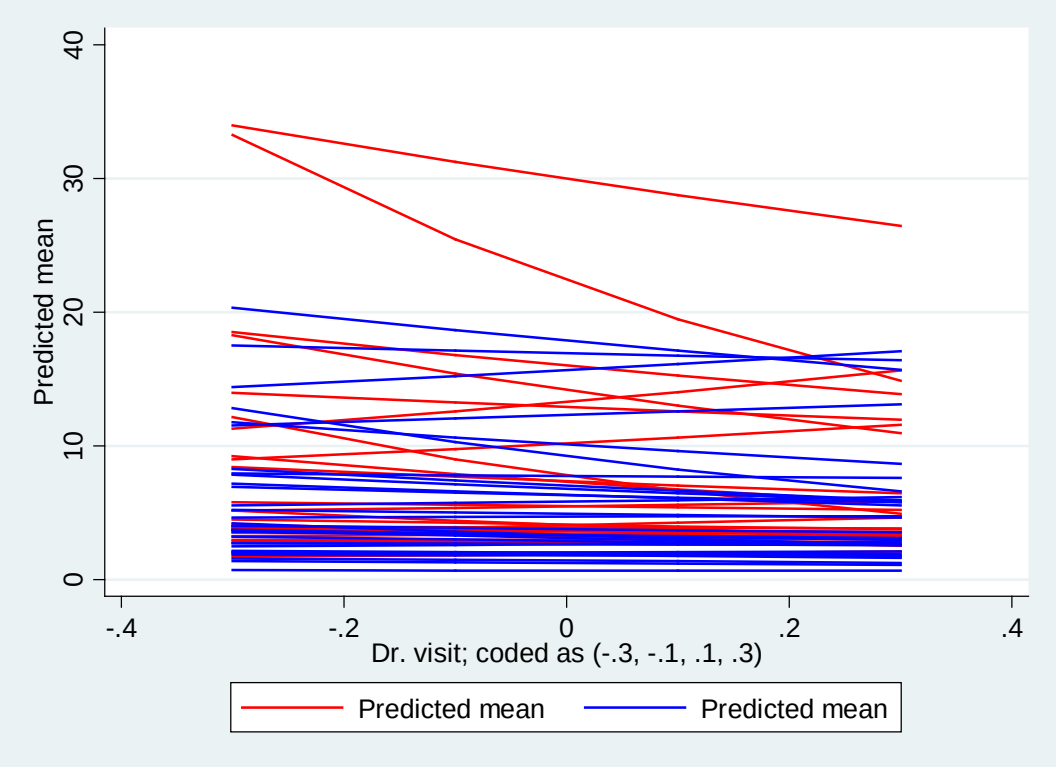
line médiaobservada médiamubeta visit if treat==0,sort(visit) lpat(dash sol-id)lcolor(red red)lwidth(thick thick)||line médiaobservada médiamubet visit if treat==1,sort(visit)lcolor(blue blue)lpat(dash solid) lwidth(thick thick)||scatter seizures visit if subject!=49&subject!=25



line médiamu médiamubeta visit if treat==0,sort(visit) lpat(dash solid)lcolor(red red)lwidth(thick thick)||line médiamu médiamubeta visit if treat==1,sort(visit)lcolor(blue blue)lpat(dash solid) lwidth(thick thick)



line mubeta visit if treat==0, sort(subject visit) con(ascending)lcolor(red)||line mubeta visit if treat==1, sort(subject visit) con(ascending)lcolor(blue)



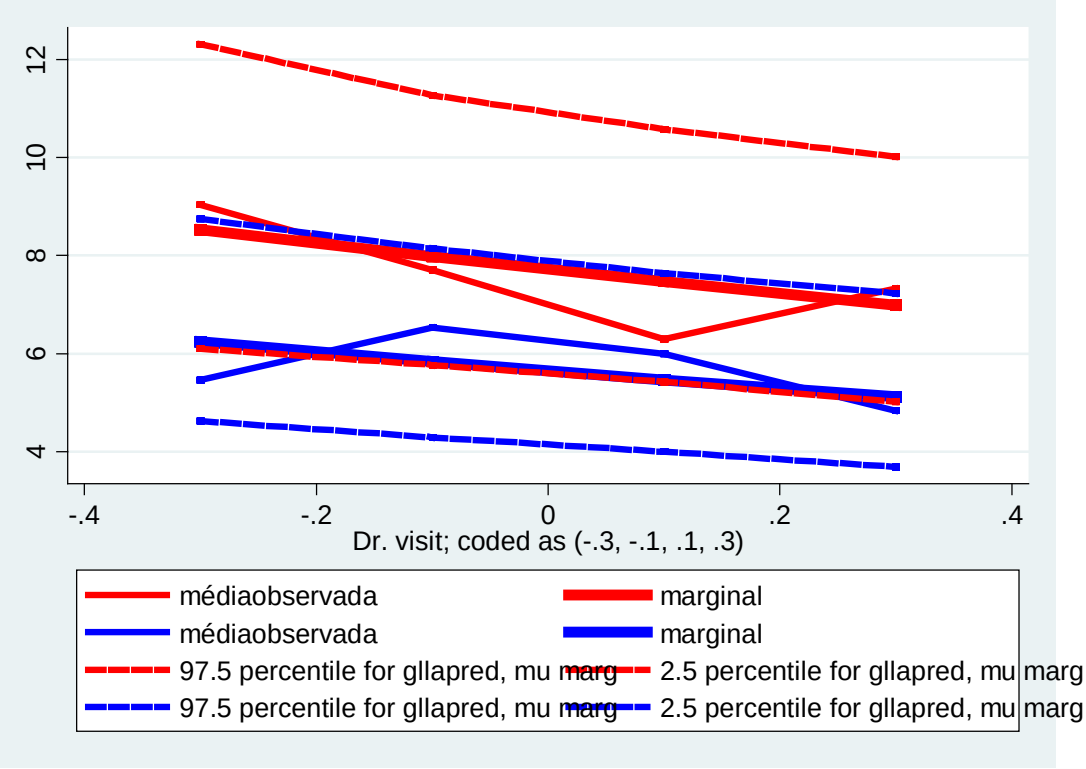
- **gllamm seizures treat visit if subject!=49 & subject!=25,i(subject) nip(15) adapt fam(poisson) link(log) eform**

```
Running adaptive quadrature
Iteration 0:    log likelihood = -648.26605
Iteration 1:    log likelihood = -628.88857
Iteration 2:    log likelihood = -628.32654
Iteration 3:    log likelihood = -628.32216
Iteration 4:    log likelihood = -628.32174


Adaptive quadrature has converged, running Newton-Raphson
Iteration 0:   log likelihood = -628.32174
Iteration 1:   log likelihood = -628.32174

number of level 1 units = 228
number of level 2 units = 57

Condition Number = 3.0675423

gllamm model

log likelihood = -628.32174

------------------------------------------------------------------------------
    seizures |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   .7318262   .1670509    -1.37   0.171     .4678512    1.144743
       visit |   .7179858   .0829225    -2.87   0.004     .5725428    .9003758
------------------------------------------------------------------------------


Variances and covariances of random effects
------------------------------------------------------------------------------


***level 2 (subject)

    var(1): .67340553 (.14384077)
------------------------------------------------------------------------------
```

- **ci_marg_mu upper lower,dots**

```
................................................
2.465 seconds = .04108333 minutes = .00068472 hours

................................................................................
> ..............................................................................
> ..............................................................................
> ..............................................................................
> ..............................................................................
> ..............................................................................
> ..............................................................................
> ..............................................................................
> ..............................................................................
> ..............................................................................
> ..............................................................................
> ..............................................................
59.592 seconds = .9932 minutes = .01655333 hours
```

line médiaobservada marginal visit if treat==0,sort(visit) lpat(solid solid)lcolor(red red)lwidth(thick vthick)||line médiaobservada marginal visit if treat==1,sort(visit)lcolor(blue blue)lpat(solid solid) lwidth(thick vthick)|| line lower visit if treat==0,sort(visit) lpat(dash)lcolor(red red)lwidth(thick thick)||line upper visit if treat==0,sort(visit)lcolor(red red)lpat(dash dash) lwidth(thick thick)|| line lower visit if treat==1,sort(visit) lpat(dash dash)lcolor(blue blue)lwidth(thick thick)||line upper visit if treat==1,sort(visit)lcolor(blue blue)lpat(dash dash) lwidth(thick thick)



5.12- Caso-controle aninhado em coorte (leva para clogit)

Dados do Apêndice VIII do livro Statistical Methods in Cancer Research, vol II, the design and analysis of cohort studies, IARC 1987, de N.E. Breslow e N.E. Day, referentes a estudo de incidência de câncer dos seios nasais em trabalhadores de metalúrgica de níquel no País de Gales.

AFE: age at first employment
BIRTHDATE: data do nascimento
EXP: quantidade de exposição
ADW: age at death or withdrawal
ASF: age at start of follow-up
ADW85: min(ADW,85); ajusta para limite de idade 85
timeexit: ADW85-AFE; tempo sob risco
timeentry: ASF-AFE; tempo sob risco, sem ser observado

escala de tempo, Barlow: stset timeexit, f(cid==160) enter(timeentry)
ou stset timeexit, f(cid==160) t0(timeentry)

escala de idade: stset ADW85,f(cid==160) origin(AFE)enter(ASF)
ou stset ADW, f(cid==160) origin(AFE)enter(ASF)

- **use "C:\HEPverãoStata\2010\nickel.dta"**
- **gen afe_10=ln(AFE-10)**
- **gen yfe_15=(BIRTHDATE+ AFE-1915)/10**
- **gen yfe_15qd= yfe_15^2**
- **gen exp= ln(EXP+1)**
- **stset ADW,f(cid==160) origin(AFE)enter(ASF) id(id)**

```
   id:  id
   failure event:  cid == 160
   obs. time interval:  (ADW[_n-1], ADW]
   enter on or after:  time ASF
   exit on or before:  failure
   t for analysis:  (time-origin)
   origin:  time AFE
------------------------------------------------------------------------------
     679  total obs.
       0  exclusions
------------------------------------------------------------------------------
         679  obs. remaining, representing
         679  subjects
          56  failures in single failure-per-subject data
 15348.06  total analysis time at risk, at risk from t =         0
                     earliest observed entry t =     9.3449
                         last observed exit t =    75.5863
```

- **set seed 123**
- **sttocc, n(3)**

```
  failure _d:  cid == 160
  analysis time _t:  (ADW-origin)
  origin:  time AFE
enter on or after:  time ASF
  id:  id

There are 56 cases
Sampling 3 controls for each case
........................................................
```

- **clogit _case afe_10 yfe_15 yfe_15qd exp, group(_set)**

```
Iteration 0:   log likelihood =  -53.16474
Iteration 1:   log likelihood = -52.798957
Iteration 2:   log likelihood = -52.798808
Iteration 3:   log likelihood = -52.798808

Conditional (fixed-effects) logistic regression      Number of obs   =        224
                                                     LR chi2(4)      =      49.67
                                                     Prob > chi2     =     0.0000
Log likelihood = -52.798808                          Pseudo R2       =     0.3199

------------------------------------------------------------------------------
       _case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      afe_10 |   1.725105   .5452455     3.16   0.002     .6564431    2.793766
      yfe_15 |   .1354012    .401618     0.34   0.736    -.6517557    .9225581
    yfe_15qd |  -1.526954   .6546366    -2.33   0.020    -2.810018   -.2438895
         exp |   .7171975   .2336319     3.07   0.002     .2592873    1.175108
------------------------------------------------------------------------------
```

- **tab _case**

```
      _case |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        168       75.00       75.00
          1 |         56       25.00      100.00
------------+-----------------------------------
      Total |        224      100.00
```

- **use “C:\HEPverãoStata\2010\nickel.dta",replace**
- **set seed 123**
- **sttocc, n(4)**

```
  failure _d:  cid == 160
  analysis time _t:  (ADW-origin)
  origin:  time AFE
  enter on or after:  time ASF
  id:  id

There are 56 cases
Sampling 4 controls for each case
.....................................................
```

- **clogit _case afe_10 yfe_15 yfe_15qd exp, group(_set)**

```
Iteration 0:   log likelihood = -59.941558
Iteration 1:   log likelihood = -59.659696
Iteration 2:   log likelihood = -59.659517
Iteration 3:   log likelihood = -59.659517

Conditional (fixed-effects) logistic regression   Number of obs   =        280
                                                  LR chi2(4)      =      60.94
                                                  Prob > chi2     =     0.0000
Log likelihood = -59.659517                       Pseudo R2       =     0.3381

------------------------------------------------------------------------------
       _case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      afe_10 |   1.751623   .5333843     3.28   0.001     .7062092    2.797037
      yfe_15 |   .0574592   .3973408     0.14   0.885    -.7213145    .8362329
    yfe_15qd |  -1.677131   .6313763    -2.66   0.008    -2.914606   -.4396563
         exp |    .755701   .2244867     3.37   0.001     .3157152    1.195687
------------------------------------------------------------------------------
```

- **tab _case**

```
      _case |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        224       80.00       80.00
          1 |         56       20.00      100.00
------------+-----------------------------------
      Total |        280      100.00
```

- **clogit _case afe_10 yfe_15 yfe_15qd exp, group(_set) or**

```
Iteration 0:   log likelihood = -59.941558
Iteration 1:   log likelihood = -59.659696
Iteration 2:   log likelihood = -59.659517
Iteration 3:   log likelihood = -59.659517

Conditional (fixed-effects) logistic regression   Number of obs   =        280
                                                  LR chi2(4)      =      60.94
                                                  Prob > chi2     =     0.0000
Log likelihood = -59.659517                       Pseudo R2       =     0.3381

------------------------------------------------------------------------------
       _case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     afe_10 |   5.763951   3.074401     3.28   0.001     2.026295    16.396
     yfe_15 |   1.059142   .4208404     0.14   0.885     .4861128    2.307657
   yfe_15qd |   .1869094   .1180102    -2.66   0.008     .0542254    .6442578
        exp |   2.129104   .4779554     3.37   0.001      1.37124    3.305828
------------------------------------------------------------------------------
```

5.13- Caso-coorte (leva para Cox)

material em STB-59, janeiro/2001,
de autoria de Vincenzo Coviello.

- **use "C:\HEPverãoStata\2010\nickel.dta"**
- **gen afe_10=ln(AFE-10)**
- **gen yfe_15=(BIRTHDATE+ AFE-1915)/10**
- **gen yfe_15qd= yfe_15^2**
- **gen exp= ln(EXP+1)**
- **stset ADW,f(cid==160) origin(AFE)enter(ASF) id(id)**

```
   id:  id
   failure event:  cid == 160
   obs. time interval:  (ADW[_n-1], ADW]
   enter on or after:  time ASF
   exit on or before:  failure
   t for analysis:  (time-origin)
   origin:  time AFE

--------------------------------------------------------------------------------
     679  total obs.
       0  exclusions
--------------------------------------------------------------------------------
         679  obs. remaining, representing
         679  subjects
          56  failures in single failure-per-subject data
15348.06  total analysis time at risk, at risk from t =         0
                              earliest observed entry t =    9.3449
                                   last observed exit t =    75.5863
```

- **stdes**

```
  failure _d:  cid == 160
  analysis time _t:  (ADW-origin)
  origin:  time AFE
enter on or after:  time ASF
  d:  id
```

Category	total	per subject mean	min	median	max
no. of subjects	679				
no. of records	679	1	1	1	1
(first) entry time		17.97722	9.3449	18.2848	36.0465
(final) exit time		40.58113	10.0427	40.0629	75.5863
subjects with gap	0				
time on gap if gap	0	.	.	.	.
time at risk	15348.057	22.60391	.4083023	21.7563	48.6659
failures	56	.0824742	0	0	1

- **stcascoh,a(.3) seed(123)**

```
  failure _d:  cid == 160
  analysis time _t:  (ADW-origin)
  origin:  time AFE
enter on or after:  time ASF
  id:  id
```

Sample composition

Subcohort member	Censored	Failure	Total
No	436	39	475
Yes	187	17	204
Total	623	56	679

Total sample = 243

No risk set with less than 4 controls

```
New stset definition

  id:  id
 failure event:  _d != 0 & _d < .
obs. time interval:  (_t0, _t]
 enter on or after:  time _t0
 exit on or before:  failure

------------------------------------------------------------------------------
     260  total obs.
       0  exclusions
------------------------------------------------------------------------------
     260  obs. remaining, representing
     243  subjects
      56  failures in single failure-per-subject data
4654.338  total analysis time at risk, at risk from t =         0
                                earliest observed entry t =  9.361301
                                     last observed exit t =   75.5863
```

- **stselpre afe_10 yfe_15 yfe_15qd exp,nohr**

```
  failure _d:  _d
  analysis time _t:  _t
 enter on or after:  time _t0
 id:  id

Method for ties: efron

Self Prentice Variance Estimate for Case-Cohort Design

Self Prentice Scheme
------------------------------------------------------------------------------------------
             |      Coef.        Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------------------
      afe_10 |   1.937362   .4888476     3.96   0.000     .9792387    2.895486
      yfe_15 |   .1818866   .3806455     0.48   0.633    -.5641649    .9279382
    yfe_15qd |  -1.477135   .6156688    -2.40   0.016    -2.683824   -.2704467
         exp |   .8238727   .2221175     3.71   0.000     .3885304    1.259215
------------------------------------------------------------------------------------------


Prentice Scheme
------------------------------------------------------------------------------------------
             |      Coef.        Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------------------
      afe_10 |    1.91569   .4888476     3.92   0.000     .9575668    2.873814
      yfe_15 |   .1561367   .3806455     0.41   0.682    -.5899148    .9021883
    yfe_15qd |  -1.430654   .6156688    -2.32   0.020    -2.637342   -.2239651
         exp |   .7999726   .2221175     3.60   0.000     .3646304    1.235315
------------------------------------------------------------------------------------------
```

- **stcox afe_10 yfe_15 yfe_15qd exp,robust nohr nolog**

```
  failure _d:  _d
  analysis time _t:  _t
  enter on or after:  time _t0
    id:  id

Cox regression -- no ties

No. of subjects      =         243              Number of obs   =       260
No. of failures      =          56
Time at risk         =  4654.338189
                                                 Wald chi2(4)    =     42.90
Log pseudolikelihood =   -220.97292              Prob > chi2     =    0.0000

                              (Std. Err. adjusted for 243 clusters in id)
------------------------------------------------------------------------------
             |               Robust
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      afe_10 |    1.91569   .4655637     4.11   0.000     1.003202    2.828178
      yfe_15 |   .1561367   .3807675     0.41   0.682    -.5901539    .9024274
   yfe_15qd |   -1.430654   .6504091    -2.20   0.028    -2.705432   -.1558754
         exp |    .7999726   .2140397     3.74   0.000     .3804625    1.219483
------------------------------------------------------------------------------
```

6. Componentes principais (*pca*), para redução de variáveis

(dados parciais da dissertação de mestrado de Daniela Wenzel, 2009)

- **use "C:\HEPverãoStata\2010\daniela.dta"**
- **des**

```
Contains data from C:\HEPverãoStata\2010\daniela.dta
  obs:           384
 vars:             6                          20 Sep 2010 15:15
 size:         9,984 (99.9% of memory free)
-------------------------------------------------------------------------------
              storage   display    value
variable name type      format     label      variable label
-------------------------------------------------------------------------------
cintura       float     %9.0g                 media cintura
diastólica    int       %9.0g                 pressao diastolica sentado braço direito
idade         float     %6.2f                 idade (ano)
imc           float     %9.0g                 IMC
pcmassagorda  float     %9.0g                 %de massa gorda
sistólica     float     %9.0g                 pressao sistolica sentado braço direito
```

- **reg diastólica cintura,noheader**

```
------------------------------------------------------------------------------
  diastólica |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      cintura |   .3830739   .0539631     7.10   0.000     .2769684    .4891795
       _cons |   45.99482   4.485378    10.25   0.000      37.1754    54.81423
------------------------------------------------------------------------------
```

- **reg diastólica imc,noheader**

```
------------------------------------------------------------------------------
  diastólica |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         imc |   1.000914   .1278446     7.83   0.000     .7495383     1.25229
       _cons |   53.28959   3.147408    16.93   0.000     47.10097    59.47821
------------------------------------------------------------------------------
```

- **reg diastólica pcmassagorda,noheader**

```
------------------------------------------------------------------------------
    diastólica |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------+----------------------------------------------------------------
pcmassagorda |   65.86378   10.33216     6.37   0.000     45.54575    86.18181
         _cons |   64.89052   2.052759    31.61   0.000      60.8538    68.92724
------------------------------------------------------------------------------
```

- **reg diastólica idade,noheader**

```
------------------------------------------------------------------------------
  diastólica |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       idade |   .3642927   .1306213     2.79   0.006     .1074572    .6211281
       _cons |   68.71627   3.244683    21.18   0.000     62.33638    75.09616
------------------------------------------------------------------------------
```

- **reg diastólica cintura imc pcmassagorda idade,noheader**

```
------------------------------------------------------------------------------
      diastólica |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------+----------------------------------------------------------------
         cintura |    .074108   .1599396     0.46   0.643    -.2404195    .3886355
             imc |   1.046474   .3602483     2.90   0.004     .3380312    1.754916
pcmassagorda |  -20.23685   24.14577    -0.84   0.403    -67.72045    27.24675
           idade |   .1055599   .1315186     0.80   0.423    -.1530765    .3641963
           _cons |   47.37602   6.708364     7.06   0.000     34.18376    60.56828
------------------------------------------------------------------------------
```

- **estat vif**

```
    Variable |       VIF       1/VIF
-------------+----------------------
     cintura |      8.58    0.116592
         imc |      7.59    0.131830
pcmassagorda |      5.70    0.175533
       idade |      1.12    0.889080
-------------+----------------------
    Mean VIF |      5.75
```

- **pca cintura imc pcmassagorda**

```
Principal components/correlation                  Number of obs    =        370
                                                  Number of comp.  =          3
                                                  Trace            =          3
Rotation: (unrotated = principal)                 Rho              =     1.0000

--------------------------------------------------------------------------
    Component |   Eigenvalue   Difference        Proportion   Cumulative
--------------+-----------------------------------------------------------
        Comp1 |      2.80105     2.67882            0.9337       0.9337
        Comp2 |      .122224    .0454926            0.0407       0.9744
        Comp3 |     .0767312           .            0.0256       1.0000
--------------------------------------------------------------------------

Principal components (eigenvectors)

--------------------------------------------------------------------------
    Variable |       Comp1       Comp2       Comp3 | Unexplained
-------------+--------------------------------------+-------------
     cintura |      0.5808      -0.3050     -0.7547 |           0
```

```
                imc |   0.5786      -0.4975       0.6463 |          0
      pcmassagorda |   0.5726       0.8121       0.1125 |          0
   ------------------------------------------------------------------------------
```

- **predict pc1scorecintimcpc,score**

(2 components skipped)

Scoring coefficients
sum of squares(column-loading) = 1

```
   ---------------------------------------------
          Variable |    Comp1     Comp2     Comp3
   ----------------+----------------------------
           cintura |   0.5808   -0.3050   -0.7547
               imc |   0.5786   -0.4975    0.6463
      pcmassagorda |   0.5726    0.8121    0.1125
   ---------------------------------------------
```

- **reg diastólica pc1scorecintimcpc ,noheader**

```
------------------------------------------------------------------------------
       diastólica |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------+-----------------------------------------------------------------
  pc1scoreci~c |   2.107218   .2859031     7.37   0.000     1.544994    2.669442
            _cons |   77.59714   .4792102   161.93   0.000     76.65478    78.5395
------------------------------------------------------------------------------
```

- **reg diastólica idade ,noheader**

```
------------------------------------------------------------------------------
 diastólica |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       idade |   .3642927   .1306213     2.79   0.006     .1074572    .6211281
       _cons |   68.71627   3.244683    21.18   0.000     62.33638    75.09616
------------------------------------------------------------------------------
```

- **reg diastólica pc1scorecintimcpc idade ,noheader**

```
------------------------------------------------------------------------------
     diastólica |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-----------------------------------------------------------------
 pc1scoreci~c |   2.031973    .299787     6.78   0.000     1.442441    2.621505
            idade |   .1094439    .130622     0.84   0.403    -.1474245    .3663124
            _cons |   74.90594   3.247542    23.07   0.000     68.51964    81.29224
------------------------------------------------------------------------------
```

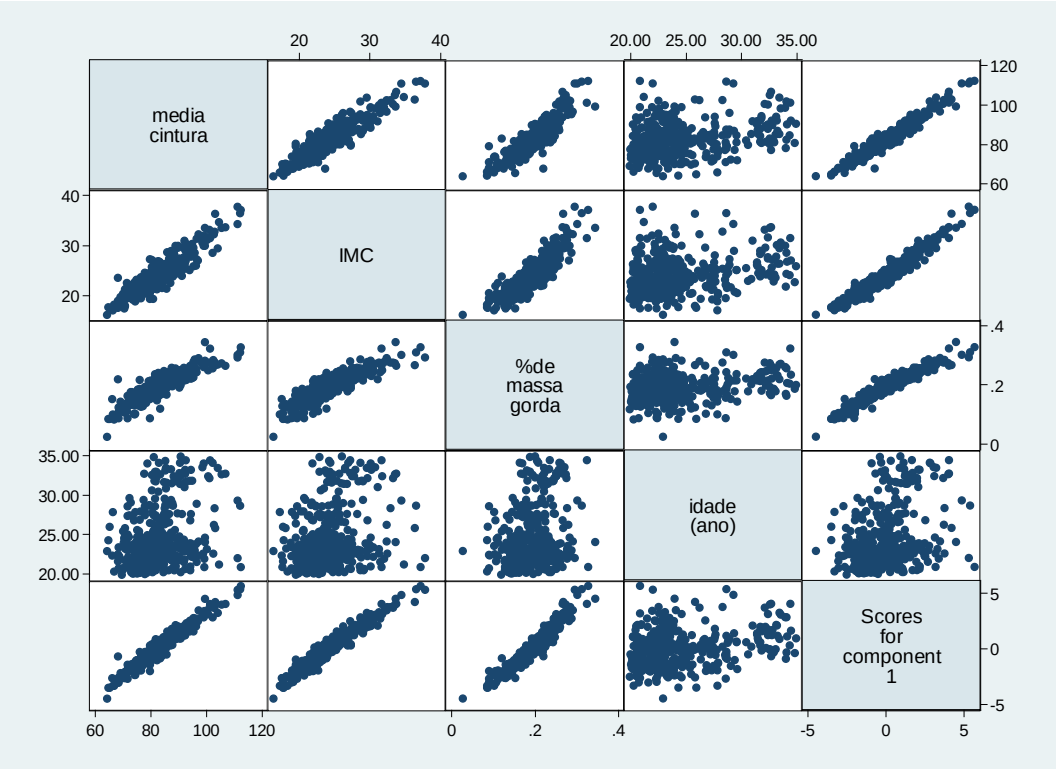
- **correlate cintura imc pcmassagorda idade pc1scorecintimcpc**

(obs=370)

```
             | cintura      imc   pcmass~a    idade   pc1sco~c
-------------+---------------------------------------------------
     cintura |  1.0000
         imc |  0.9224   1.0000
pcmassagorda |  0.8948   0.8842   1.0000
       idade |  0.3269   0.2829   0.2625   1.0000
```

pc1scoreci~c | 0.9721 0.9683 0.9583 0.3010 1.0000

- **graph matrix cintura imc pcmassagorda idade pc1scorecintimcpc**



7. Tamanho de amostra

7.1- Teste de hipóteses

Teste t de Student para detectar possível diferença entre média1=132,86 vs média2= 127,44, com desvio padrão1= 15,34 e desvio padrão2= 18,23, poder do teste= 80%, nível de significância= 5% bicaudal, tamanho da amostra n2= dobro de n1.

- **sampsi 132.86 127.44,p(.8) r(2) sd1(15.34) sd2(18.23)**

```
Estimated sample size for two-sample comparison of means
Test Ho: m1 = m2, where m1 is the mean in population 1 and m2 is the mean in population 2
Assumptions:
        alpha =   0.0500  (two-sided)
        power =   0.8000
           m1 =   132.86
           m2 =   127.44
          sd1 =    15.34
          sd2 =    18.23
        n2/n1 =     2.00
Estimated required sample sizes:
           n1 =      108
           n2 =      216
```

Mesma situação, mas conhecer o poder se só conseguir n1=n2=100

- **sampsi 132.86 127.44,n1=100 sd1(15.34) sd2(18.23)**

```
Estimated power for two-sample comparison of means
Test Ho: m1 = m2, where m1 is the mean in population 1  and m2 is the mean in population 2
Assumptions:
        alpha =   0.0500  (two-sided)
           m1 =   132.86
           m2 =   127.44
          sd1 =    15.34
          sd2 =    18.23
      sample size n1 =      100
                  n2 =      100
              n2/n1 =     1.00
Estimated power:
        power =   0.6236
```

Obter o poder do teste ao comparar duas proporções 25% vs 40%, com tamanhos de amostras n1=300 e n2= metade de n1.

- **sampsi .25 .4, n1(300) r(.5)**

```
Estimated power for two-sample comparison of proportions
Test Ho: p1 = p2, where p1 is the proportion in population 1  and p2 is the proportion in population 2
Assumptions:
        alpha =   0.0500  (two-sided)
           p1 =   0.2500
           p2 =   0.4000
sample size n1 =      300
           n2 =      150
           n2/n1 =     0.50
Estimated power:
```

power = 0.8790

Exemplo de cálculo de tamanho de amostra para comparar proporções

aleitamento	deficiência de vit. A: não	sim	tamanho da amostra
sim (1)		p1= 25%	100% n1= ?
não (2)		p2= 40%	100% n2= ?

Imagine que seja conhecido que 25% (p1= .25) das crianças com aleitamento materno nos cinco primeiros meses de idade fiquem com deficiência de vitamina A.
Uma hipótese com base em vários conhecimentos fisiológicos, biológicos, clínicos, epidemiológicos é que, em crianças sem aleitamento, esta prevalência deve ser maior, possivelmente uns 40% (p2= .40).
Quais os tamanhos das amostras (n1 e n2, sendo n1=n2) para detectar estatisticamente esta diferença, com um nível de 5% de significância e poder do teste de 80%?

- sampsi .25 .40, power(.80) a(.05)

```
Estimated sample size for two-sample comparison of proportions
Test Ho: p1 = p2, where p1 is the proportion in population 1
                  and p2 is the proportion in population 2
Assumptions:
         alpha =   0.0500  (two-sided)
         power =   0.8000
            p1 =   0.2500
            p2 =   0.4000
         n2/n1 =   1.00
Estimated required sample sizes:
            n1 =      165
            n2 =      165
```

Amostra total= 165+165= 330

Mas se for possível conseguir somente n1= 100, qual o tamanho da amostra de crianças sem aleitamento para compensar, mantendo os mesmos 5% de significância e 80% de poder?

((1/165)+(1/165))=((1/100)+(1/n2))
2/165 = (1/100)+(1/n2)
n2= 1/((2/165)-(1/100))

- dis 1/((2/165)-(1/100))

471.42857 ⟶ n2= 472 Amostra total= 100+472= 572

- dis (1/100)+(1/472)

.01212

- dis (1/165)+(1/165)

```
.01212
```

Estudo caso-controle

lembrar a tabela de resultados:

EXPOSIÇÃO AO FATOR	CONDIÇÃO Caso (1)	Controle (2)	
Presente (1)	a (p1)	b (p2)	m_1
Ausente (0)	c (1-p1)	d (1-p2)	m_0
	n_1 (1=100%)	n_2 (1=100%)	T

"ODDS" da presença de exposição entre os casos: $\frac{a}{c}$ **= p1/(1-p1)**

"ODDS" da presença de exposição entre os controles: $\frac{b}{d}$ **=p2/(1-p2)**

"ODDS RATIO": $\frac{a/c}{b/d} = \frac{a.d}{b.c}$ **= [p1x(1-p2)]/[p2x(1-p1)]=** $\wedge$ **OR**

$$p_2 = \frac{p_1}{or(1-p_1)+p_1} = \frac{p_1}{or + p_1(1-or)}$$

$$p_1 = \frac{or.p_2}{1+p_2(or-1)}$$

Tamanhos (iguais) das amostras n1 e n2 para detectar odds ratio mínimo 1,5, nível 5% de significância bicaudal, poder do teste 80%, com 39% dos casos expostos ao fator de risco (p1):

- **dis .39/(1.5+(.39*(1-1.5)))**

  p2= .29885057= .30

- **sampsi .39 .30, p(.8)**

```
Estimated sample size for two-sample comparison of proportions
Test Ho: p1 = p2, where p1 is the proportion in population 1
        and p2 is the proportion in population 2
Assumptions:
     alpha =  0.0500  (two-sided)
     power =  0.8000
        p1 =  0.3900
        p2 =  0.3000
     n2/n1 =  1.00
Estimated required sample sizes:
        n1 =     459
        n2 =     459
```

Qual o tamanho n2 se for possível obter somente 300 casos?

1/460 + 1/460= 1/300 + 1/n2
1/n2= 1/460 +1/460 – 1/300
n2= 1/(1/460 +1/460 – 1/300)

- **dis 1/(1/460 +1/460 - 1/300)**

  985.71429= 986

- **sampsi .39 .30, n1(300)n2(986)alpha(.05)**

Estimated power for two-sample comparison of proportions
Test Ho: p1 = p2, where p1 is the proportion in population 1
and p2 is the proportion in population 2
Assumptions:
alpha = 0.0500 (two-sided)
p1 = 0.3900
p2 = 0.3000
sample size n1 = 300
n2 = 986
n2/n1 = 3.29
Estimated power:
power = 0.8070

7.2 - Intervalo de confiança

$$\bar{x} - z_{\alpha/2}.\sigma_{\bar{x}} \frac{\sigma}{\sqrt{n}} \overset{A}{\longleftrightarrow} \bar{x} + z_{\alpha/2}.\sigma_{\bar{x}}$$

$$\bar{x} - z_{\alpha/2}.\frac{\sigma}{\sqrt{n}} \overset{A}{\longleftrightarrow} \bar{x} + z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}$$

$$A = 2.z_{\alpha/2}.\frac{\sigma}{\sqrt{n}} \Rightarrow n = \frac{4.z^2_{\alpha/2}.\sigma^2}{A^2}$$

Média em torno de 130g, desvio padrão aproximado= 15g α=5%, A= 15g:

- **display (4*1.96^2*15^2)/ 15^2**

  15.3664 → 16

-

- **dis 130-1.96*15/sqrt(16)**

  122.65

-

- **dis 130+1.96*15/sqrt(16)**

137.35

Proporção p= 0,5, que leva para o maior tamanho de amostra, A= 0,1:

- **display (4* 1.96^2*.25)/.1^2**

  384.16 → 400

- **dis .5-(1.96*sqrt(.25/400))**

  .45

- **dis .5+(1.96*sqrt(.25/400))**

  .55

7.3 – Outros comandos úteis, a partir de *findit* ou *ssc install*

fpower

simpower

powerreg

sampsi_reg

samplesize

xsampsi

studysi

8. *fweight, freq; expand*

Construir o banco lele.dta, usando o editor:

```
escol  fumar peso contagem
      0      1     1        45
      1      1     1       132
      0      0     0       886
      0      0     1        55
      0      1     0       422
      1      1     0       943
      1      0     1       140
      1      0     0      1626
```

- **tab peso fumar if escol==0,chi col**

```
           |         fumar
      peso |         0          1 |     Total
-----------+----------------------+----------
         0 |         1          1 |         2
           |     50.00      50.00 |     50.00
-----------+----------------------+----------
         1 |         1          1 |         2
           |     50.00      50.00 |     50.00
-----------+----------------------+----------
     Total |         2          2 |         4
           |    100.00     100.00 |    100.00

          Pearson chi2(1) =   0.0000   Pr = 1.000
```

- **tab peso fumar if escol==1,chi col**

```
           |         fumar
      peso |         0          1 |     Total
-----------+----------------------+----------
         0 |         1          1 |         2
           |     50.00      50.00 |     50.00
-----------+----------------------+----------
         1 |         1          1 |         2
           |     50.00      50.00 |     50.00
-----------+----------------------+----------
     Total |         2          2 |         4
           |    100.00     100.00 |    100.00

          Pearson chi2(1) =   0.0000   Pr = 1.000
```

- **tab peso fumar ,chi col**

```
           |         fumar
      peso |         0          1 |     Total
-----------+----------------------+----------
         0 |         2          2 |         4
           |     50.00      50.00 |     50.00
-----------+----------------------+----------
         1 |         2          2 |         4
           |     50.00      50.00 |     50.00
-----------+----------------------+----------
     Total |         4          4 |         8
           |    100.00     100.00 |    100.00

          Pearson chi2(1) =   0.0000   Pr = 1.000
```

- **cc peso fumar, by(escol) woolf**

```
           escol |       OR       [95% Conf. Interval]   M-H Weight
-----------------+-------------------------------------------------
               0 |          1      .0198425   50.39681          .25 (Woolf)
               1 |          1      .0198425   50.39681          .25 (Woolf)
-----------------+-------------------------------------------------
           Crude |          1      .0625488   15.98751              (Woolf)
    M-H combined |          1      .0625488   15.98751
-----------------------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     0.00  Pr>chi2 = 1.0000

                   Test that combined OR = 1:
                               Mantel-Haenszel chi2(1) =      0.00
                                                Pr>chi2 =    1.0000
```

- **tab peso fumar [freq= contagem] if escol==0,chi col**

```
           |         fumar
      peso |         0          1 |     Total
-----------+----------------------+----------
         0 |       886        422 |     1,308
           |     94.16      90.36 |     92.90
-----------+----------------------+----------
         1 |        55         45 |       100
           |      5.84       9.64 |      7.10
-----------+----------------------+----------
     Total |       941        467 |     1,408
           |    100.00     100.00 |    100.00

          Pearson chi2(1) =   6.7989   Pr = 0.009
```

- **tab peso fumar [freq= contagem] if escol==1,chi col**

```
           |         fumar
      peso |         0          1 |     Total
-----------+----------------------+----------
         0 |     1,626        943 |     2,569
           |     92.07      87.72 |     90.43
-----------+----------------------+----------
         1 |       140        132 |       272
           |      7.93      12.28 |      9.57
-----------+----------------------+----------
     Total |     1,766      1,075 |     2,841
           |    100.00     100.00 |    100.00

          Pearson chi2(1) =  14.6159   Pr = 0.000
```

- **tab peso fumar [freq= contagem] ,chi col**

```
           |         fumar
      peso |         0          1 |     Total
-----------+----------------------+----------
         0 |     2,512      1,365 |     3,877
           |     92.80      88.52 |     91.24
-----------+----------------------+----------
         1 |       195        177 |       372
           |      7.20      11.48 |      8.76
-----------+----------------------+----------
     Total |     2,707      1,542 |     4,249
           |    100.00     100.00 |    100.00

          Pearson chi2(1) =  22.4752   Pr = 0.000
```

- **cc peso fumar [freq= contagem], by(escol) woolf**

```
           escol |       OR       [95% Conf. Interval]   M-H Weight
-----------------+-------------------------------------------------
               0 |   1.717794      1.139252   2.590135     16.48438 (Woolf)
               1 |   1.625754      1.264933   2.089498     46.46955 (Woolf)
-----------------+-------------------------------------------------
           Crude |    1.67042      1.348767    2.06878              (Woolf)
    M-H combined |   1.649854      1.331762   2.043924
-----------------------------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     0.05  Pr>chi2 = 0.8225

                  Test that combined OR = 1:
                              Mantel-Haenszel chi2(1) =     21.33
                                                Pr>chi2 =    0.0000
```

.

- **logistic peso escol fumar**

```
Logistic regression                               Number of obs   =          8
                                                  LR chi2(2)      =       0.00
                                                  Prob > chi2     =     1.0000
Log likelihood = -5.5451774                       Pseudo R2       =     0.0000

------------------------------------------------------------------------------
       peso | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
      escol |          1   1.414214     0.00   1.000     .0625488    15.98751
      fumar |          1   1.414214     0.00   1.000     .0625488    15.98751
------------------------------------------------------------------------------
```

- **logistic peso escol fumar [freq= contagem]**

```
Logistic regression                               Number of obs   =       4249
                                                  LR chi2(2)      =      28.15
                                                  Prob > chi2     =     0.0000
Log likelihood = -1247.1643                       Pseudo R2       =     0.0112

------------------------------------------------------------------------------
       peso | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
      escol |   1.353334   .1653111     2.48   0.013     1.065197    1.719412
      fumar |   1.650245   .1803683     4.58   0.000      1.33203     2.04448
------------------------------------------------------------------------------
```

- **expand contagem**

```
(4241 observations created)

. list in 1/15, clean

      escol   fumar   peso   contagem
  1.      0       1      0        422
  2.      1       1      1        132
  3.      0       0      0        886
  4.      0       1      1         45
  5.      0       0      1         55
  6.      1       0      1        140
  7.      1       0      0       1626
  8.      1       1      0        943
  9.      0       1      0        422
 10.      0       1      0        422
 11.      0       1      0        422
 12.      0       1      0        422
 13.      0       1      0        422
 14.      0       1      0        422
 15.      0       1      0        422
```

- **logistic peso escol fumar**

```
Logistic regression                               Number of obs   =       4249
                                                  LR chi2(2)      =      28.15
                                                  Prob > chi2     =     0.0000
Log likelihood = -1247.1643                       Pseudo R2       =     0.0112

------------------------------------------------------------------------------
        peso | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       escol |   1.353334   .1653111     2.48   0.013     1.065197    1.719412
       fumar |   1.650245   .1803683     4.58   0.000      1.33203     2.04448
------------------------------------------------------------------------------
```

9. *svy* (amostragem complexa) (Colaboração de Rafael Claro e de Renata B. Levy)

POPULAÇÃO DE ESTUDO, COM DOIS GRANDES GRUPOS: grupo 0 e grupo 1.
HÁ UM FATOR (EXPOSIÇÃO) QUE PODE CAUSAR PROBLEMA DE SAÚDE. CADA INDIVÍDUO PODE ESTAR exposto (=1) ou não exposto (=0).
APÓS ALGUM TEMPO FIXO DE EXPOSIÇÃO OU NÃO EXPOSIÇÃO, CADA INDIVÍDUO PODE TER (resposta= 1) OU NÃO TER PROBLEMA DE SAÚDE (resposta = 0)

- **use "C:\HEPverãoStata\2010\amostracomplexa.dta"**
- **tab resp exp [freq= contagem] if grupo==0, col**

```
           |          exp
      resp |         0          1 |     Total
-----------+----------------------+----------
         0 |     1,080        630 |     1,710
           |     80.00      70.00 |     76.00
-----------+----------------------+----------
         1 |       270        270 |       540
           |     20.00      30.00 |     24.00
-----------+----------------------+----------
     Total |     1,350        900 |     2,250
           |    100.00     100.00 |    100.00
```

```
. dis 270*1080/630/270
1.7142857
. dis (270/900)/(270/1350)
1.5
```

- **tab resp exp [freq= contagem] if grupo==1, col**

```
           |          exp
      resp |         0          1 |     Total
-----------+----------------------+----------
         0 |        30         80 |       110
           |     60.00      40.00 |     44.00
-----------+----------------------+----------
         1 |        20        120 |       140
           |     40.00      60.00 |     56.00
-----------+----------------------+----------
     Total |        50        200 |       250
           |    100.00     100.00 |    100.00
```

```
. dis 120*30/80/20
2.25
. dis (120/200)/(20/50)
1.5
```

- **tab resp exp [freq= contagem] , col**

```
           |          exp
      resp |         0          1 |     Total
-----------+----------------------+----------
         0 |     1,110        710 |     1,820
           |     79.29      64.55 |     72.80
-----------+----------------------+----------
         1 |       290        390 |       680
           |     20.71      35.45 |     27.20
-----------+----------------------+----------
     Total |     1,400      1,100 |     2,500
           |    100.00     100.00 |    100.00
```

```
. dis 390*1110/710/290
2.1024769
. dis (390/1100)/(290/1400)
```

1.7115987

- **cc resp exp [freq= contagem], by(grupo)**

```
          grupo |      OR      [95% Conf. Interval]   M-H Weight
----------------+--------------------------------------------------
              0 |  1.714286     1.403492   2.093943        75.6 (exact)
              1 |      2.25     1.14277    4.480475        6.4 (exact)
----------------+--------------------------------------------------
          Crude |  2.102477     1.751181   2.524421             (exact)
M-H combined |  1.756098     1.457459   2.115929
---------------------------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     0.65  Pr>chi2 = 0.4208
 Test that combined OR = 1:
                        Mantel-Haenszel chi2(1) =     35.45
                                 Pr>chi2 =    0.0000
```

- **cs resp exp [freq= contagem], by(grupo)**

```
          grupo |      RR      [95% Conf. Interval]   M-H Weight
----------------+--------------------------------------------------
              0 |       1.5     1.296124   1.735945         108
              1 |       1.5     1.048779   2.145353          16
----------------+--------------------------------------------------
          Crude |   1.711599     1.503177   1.948918
M-H combined |       1.5     1.310108   1.717416
---------------------------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =    0.000  Pr>chi2 = 1.0000
```

- **logistic resp exp [freq= contagem] if grupo==0**

```
Logistic regression                             Number of obs   =       2250
                                                LR chi2(1)      =      29.22
                                                Prob > chi2     =     0.0000
Log likelihood = -1225.3211                     Pseudo R2       =     0.0118
------------------------------------------------------------------------------
        resp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   1.714286   .1707435     5.41   0.000     1.410273    2.083834
------------------------------------------------------------------------------
```

- **oddsrisk resp exp [freq= contagem] if grupo==0**

```
---------------------------------------------------------------------
Incidence for unexposed risk group =     0.2000
---------------------------------------------------------------------
Predictor    Odds Ratio   Risk Ratio     [95% Conf. Interval]
---------------------------------------------------------------------
exp                1.7143       1.5000          1.3033       1.7126
---------------------------------------------------------------------
```

- **logistic resp exp [freq= contagem] if grupo==1**

```
Logistic regression                     Number of obs   =        250
                                        LR chi2(1)      =       6.46
                                        Prob > chi2     =     0.0110
Log likelihood = -168.25292             Pseudo R2       =     0.0188
------------------------------------------------------------------------------
        resp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |        2.25   .7261843     2.51   0.012     1.195249    4.235519
------------------------------------------------------------------------------
```

- **oddsrisk resp exp [freq= contagem] if grupo==1**

```
------------------------------------------------------------------------------
Incidence for unexposed risk group =     0.4000
------------------------------------------------------------------------------
Predictor    Odds Ratio   Risk Ratio    [95% Conf. Interval]
------------------------------------------------------------------------------
exp              2.2500       1.5000      1.1087      1.8462
------------------------------------------------------------------------------
```

- **logistic resp exp [freq= contagem]**

```
Logistic regression                     Number of obs   =       2500
                                        LR chi2(1)      =      67.31
                                        Prob > chi2     =     0.0000
Log likelihood = -1429.4417             Pseudo R2       =     0.0230
------------------------------------------------------------------------------
        resp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   2.102477   .1917953     8.15   0.000     1.758254     2.51409
------------------------------------------------------------------------------
```

- **oddsrisk resp exp [freq= contagem]**

```
------------------------------------------------------------------------------
Incidence for unexposed risk group =     0.2071
------------------------------------------------------------------------------
Predictor    Odds Ratio   Risk Ratio    [95% Conf. Interval]
------------------------------------------------------------------------------
exp              2.1025       1.7116      1.5196      1.9138
------------------------------------------------------------------------------
```

- **logistic resp exp grupo [freq= contagem]**

```
Logistic regression                     Number of obs   =       2500
                                        LR chi2(2)      =     138.39
                                        Prob > chi2     =     0.0000
Log likelihood = -1393.9011             Pseudo R2       =     0.0473
------------------------------------------------------------------------------
        resp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   1.755859   .1669562     5.92   0.000     1.457313    2.115564
       grupo |   3.286924   .4634589     8.44   0.000     2.493271     4.33321
------------------------------------------------------------------------------
```

- **oddsrisk resp exp grupo [freq= contagem]**

```
------------------------------------------------------------------------------
Incidence for unexposed risk group =     0.2071
------------------------------------------------------------------------------
Predictor    Odds Ratio   Risk Ratio    [95% Conf. Interval]
------------------------------------------------------------------------------
exp              1.7559       1.5182      1.3312      1.7185
grupo            3.2869       2.2304      1.9042      2.5633
------------------------------------------------------------------------------
```

Usando amostragem casual simples, com n/N= 500/2500= 20%

- **use "C:\HEPverãoStata\2010\amostracomplexaproporcional.dta", clear**
- **bysort estrato:tab resp exp, col**

```
-> estrato = 0

           |          exp
      resp |         0          1 |     Total
-----------+----------------------+----------
         0 |       216        126 |       342
           |     80.00      70.00 |     76.00
-----------+----------------------+----------
         1 |        54         54 |       108
           |     20.00      30.00 |     24.00
-----------+----------------------+----------
     Total |       270        180 |       450
           |    100.00     100.00 |    100.00

-> estrato = 1
           |          exp
      resp |         0          1 |     Total
-----------+----------------------+----------
         0 |         6         16 |        22
           |     60.00      40.00 |     44.00
-----------+----------------------+----------
         1 |         4         24 |        28
           |     40.00      60.00 |     56.00
-----------+----------------------+----------
     Total |        10         40 |        50
           |    100.00     100.00 |    100.00
```

- **tab resp exp, col**

```
           |          exp
      resp |         0          1 |     Total
-----------+----------------------+----------
         0 |       222        142 |       364
           |     79.29      64.55 |     72.80
-----------+----------------------+----------
         1 |        58         78 |       136
           |     20.71      35.45 |     27.20
-----------+----------------------+----------
     Total |       280        220 |       500
           |    100.00     100.00 |    100.00
```

- **logistic resp exp estrato**

```
Logistic regression                               Number of obs   =        500
                                                  LR chi2(2)      =      27.68
                                                  Prob > chi2     =     0.0000
Log likelihood = -278.78022                       Pseudo R2       =     0.0473

------------------------------------------------------------------------------
        resp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   1.755859   .3733253     2.65   0.008      1.15747    2.663603
     estrato |   3.286924   1.036326     3.77   0.000     1.771808    6.097654
------------------------------------------------------------------------------
```

- **cc resp exp,by(estrato)**

```
         estrato |       OR       [95% Conf. Interval]   M-H Weight
-----------------+------------------------------------------------
               0 |   1.714286      1.080962   2.715003        15.12 (exact)
               1 |       2.25      .4428824   12.46726         1.28 (exact)
-----------------+------------------------------------------------
           Crude |   2.102477      1.382516   3.201636              (exact)
    M-H combined |   1.756098      1.157533   2.664181
-----------------------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     0.13  Pr>chi2 = 0.7188
                  Test that combined OR = 1:
                              Mantel-Haenszel chi2(1) =      7.07
                                              Pr>chi2 =    0.0078
```

- **cs resp exp,by(estrato)**

```
         estrato |       RR       [95% Conf. Interval]   M-H Weight
-----------------+------------------------------------------------
               0 |         1.5     1.081993   2.079496         21.6
               1 |         1.5      .673891   3.338819          3.2
-----------------+------------------------------------------------
           Crude |    1.711599     1.280284   2.288219
    M-H combined |         1.5      1.10827   2.030191
-----------------------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =    0.000  Pr>chi2 = 1.0000

.
```

- **use "C:\HEPverãoStata\2010\amostracomplexasvy.dta", clear**

Agora a amostragem é "complexa". Todos os membros do grupo 1 pertencem à amostra; portanto, cada um deles entrou com probabilidade= 1. Os membros do grupo 0 entraram com probabilidade 0,20 cada um (450/2250).

Define-se uma quantidade pweight= 1/probabilidade, que pode ser interpretada como a quantidade de membros na população que cada indivíduo da amostra representa. Se pweight= 1, na população haverá 1 indivíduo com as mesmas características desta pessoa da amostra. Se o pweight de um indivíduo da amostra for pweight= 1/probabilidade= 1/0,20= 5, então haverá 5 indivíduos iguais a este na população de estudo.

- **tab resp exp if grupo==0, col**

resp	exp 0	exp 1	Total
0	216	126	342
	80.00	70.00	76.00
1	54	54	108
	20.00	30.00	24.00
Total	270	180	450
	100.00	100.00	100.00

- **tab resp exp if grupo==1, col**

resp	exp 0	exp 1	Total
0	30	80	110
	60.00	40.00	44.00
1	20	120	140
	40.00	60.00	56.00
Total	50	200	250
	100.00	100.00	100.00

- **svyset [pweight= pw]**

```
  pweight: pw
      VCE: linearized
Single unit: missing
   Strata 1: <one>
      SU 1: <observations>
     FPC 1: <zero>
```

- **logistic resp exp**

```
Logistic regression                             Number of obs   =        700
                                                LR chi2(1)      =      39.89
                                                Prob > chi2     =     0.0000
Log likelihood = -435.09731                     Pseudo R2       =     0.0438
------------------------------------------------------------------------------
        resp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   2.807924   .4713638     6.15   0.000     2.020668    3.901898
------------------------------------------------------------------------------
```

- **logistic resp grupo exp**

```
Logistic regression                             Number of obs   =        700
                                                LR chi2(2)      =      82.97
                                                Prob > chi2     =     0.0000
Log likelihood = -413.55879                     Pseudo R2       =     0.0912
------------------------------------------------------------------------------
        resp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       grupo |   3.219095    .578677     6.50   0.000     2.263177    4.578773
         exp |   1.872805   .3424205     3.43   0.001     1.308761    2.679939
------------------------------------------------------------------------------
```

- **logistic resp exp if grupo==0**

```
Logistic regression                             Number of obs   =        450
                                                LR chi2(1)      =       5.84
                                                Prob > chi2     =     0.0156
Log likelihood = -245.06423                     Pseudo R2       =     0.0118
------------------------------------------------------------------------------
        resp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   1.714286   .3817941     2.42   0.016     1.107922     2.65251
------------------------------------------------------------------------------
```

- **logistic resp exp if grupo==1**

```
Logistic regression                             Number of obs   =        250
                                                LR chi2(1)      =       6.46
                                                Prob > chi2     =     0.0110
Log likelihood = -168.25292                     Pseudo R2       =     0.0188
------------------------------------------------------------------------------
        resp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |       2.25   .7261843     2.51   0.012     1.195249    4.235519
------------------------------------------------------------------------------
```

- **svy: logistic resp exp**

```
(running logistic on estimation sample)
Survey: Logistic regression
Number of strata   =         1              Number of obs      =       700
Number of PSUs     =       700              Population size    =      2500
                                            Design df          =       699
                                            F(   1,    699)    =     15.00
                                            Prob > F           =    0.0001
------------------------------------------------------------------------------
             |              Linearized
        resp | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   2.102477   .4034455     3.87   0.000     1.442482    3.064447
------------------------------------------------------------------------------
```

- **svy: logistic resp grupo exp**

```
(running logistic on estimation sample)
Survey: Logistic regression
Number of strata   =         1              Number of obs      =       700
Number of PSUs     =       700              Population size    =      2500
                                            Design df          =       699
                                            F(   2,    698)    =     36.54
                                            Prob > F           =    0.0000
------------------------------------------------------------------------------
             |              Linearized
        resp | Odds Ratio   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       grupo |   3.286924   .5977963     6.54   0.000     2.299917    4.697503
         exp |   1.755859   .3600672     2.75   0.006     1.173908    2.626304
------------------------------------------------------------------------------
```

- **tab resp exp,col**

```
           |        exp
      resp |         0          1 |     Total
-----------+----------------------+----------
         0 |       246        206 |       452
           |     76.88      54.21 |     64.57
-----------+----------------------+----------
         1 |        74        174 |       248
           |     23.13      45.79 |     35.43
-----------+----------------------+----------
     Total |       320        380 |       700
           |    100.00     100.00 |    100.00
```

- **svy:tab resp exp,col**

```
(running tabulate on estimation sample)
Number of strata   =         1              Number of obs      =       700
Number of PSUs     =       700              Population size    =      2500
                                            Design df          =       699
-------------------------------
           |       exp
      resp |    0        1       Total
-----------+-------------------
         0 | .7929     .6455    .728
         1 | .2071     .3545    .272
           |
     Total |    1          1         1
-------------------------------
Key:  column proportions
Pearson:
  Uncorrected   chi2(1)        =   18.9256
  Design-based  F(1, 699)      =   15.2756     P = 0.0001
```

- **tab resp exp,col chi**

```
           |        exp
      resp |        0         1 |     Total
-----------+----------------------+----------
         0 |      246       206 |       452
           |    76.88     54.21 |     64.57
-----------+----------------------+----------
         1 |       74       174 |       248
           |    23.13     45.79 |     35.43
-----------+----------------------+----------
     Total |      320       380 |       700
           |   100.00    100.00 |    100.00
  Pearson chi2(1) =  39.0061   Pr = 0.000
```

- **svy: tab resp exp ,col count**

```
(running tabulate on estimation sample)
Number of strata   =         1              Number of obs      =       700
Number of PSUs     =       700              Population size    =      2500
                                            Design df          =       699
 --------------------------------
           |        exp
      resp |    0         1    Total
-----------+--------------------
         0 | 1110       710     1820
           | .7929    . 6455    .728
           |
         1 |  290       390      680
           | .2071      .3545    .272
           |
     Total | 1400      1100     2500
           |    1         1        1
 --------------------------------
 Key:  weighted counts
       column proportions
 Pearson:
   Uncorrected   chi2(1)        =   18.9256
   Design-based  F(1, 699)      =   15.2756    P = 0.0001
```

10. Arquivos *.do; breve introdução

Às vezes é necessário realizar uma análise igual para conjuntos de dados diferentes. Isto é possível, armazenando-se os comandos em um arquivo com extensão **.do**.

Uma forma de criar um arquivo *.do é salvando os comandos utilizados durante a sessão de trabalho. Isto pode ser feito selecionando usando a sequência: Apontar o mouse para Command, clicar com o botão da direita, clicar "select all". Clicar outra vez com o botão da direita, agora em "Send to Do-file Editor". Salvar com o nome de preferência (a extensão automaticamente será .do) em pasta adequada.
Qualquer processador de texto pode ser utilizado para a correção dos comandos, lembrando que o arquivo *.do é texto, em ASCII. A seguir é apresentada uma estrutura básica de um arquivo *.do:

```
*comentário descrevendo o que o arquivo faz*
capture log close
log using filename, append
set more off
command 1
command 2
.
.
log
close
exit
```

Cada linha significa:

1. Os asteriscos fazem com que seja ignorado o que está entre eles; são usados para comentários.
2. O comando **capture log close** fecha o arquivo **log** em uso se houver uma mensagem de erro.
3. O comando **log using filename, append** abre um arquivo **log** que salvará os resultados abaixo de um já existente.
4. O comando **set more off** faz com que a saída seja apresentada na tela automaticamente, sem ter que manualmente instruir o *Stata* para mostrar o que está faltando.
5. Depois que a lista de comandos já estiver digitada e os resultados prontos, o arquivo **.log** é fechado com o comando **log close**.
6. A última linha do programa contendo o comando **exit** faz com que o programa pare de ser rodado.

Para abrir um arquivo *.do, pressionar com o *mouse* o ícone à esquerda do Editor.

Digitar a seqüência de comandos:

```
*Analise de baixo peso ao nascer por sexo*
capture log close
log using c:\ HEPStata/VERÃO/2013/docampinas.do
set more off
use c:/HEPStata/VERÃO/2013/campinas.dta,clear
mvdecode  _all, mv(-99)
sum pesonasc, d
gen baixope=1 if pesonasc<2500
replace baixope=0 if pesonasc>=2500
tab baixope sexo, chi2 row col
log close
```

exit

Após o término da digitação, salvar com: **File, Save as docampinas.do**

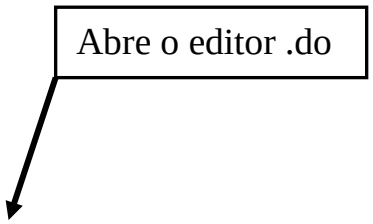
Para executar o programa, digitar:

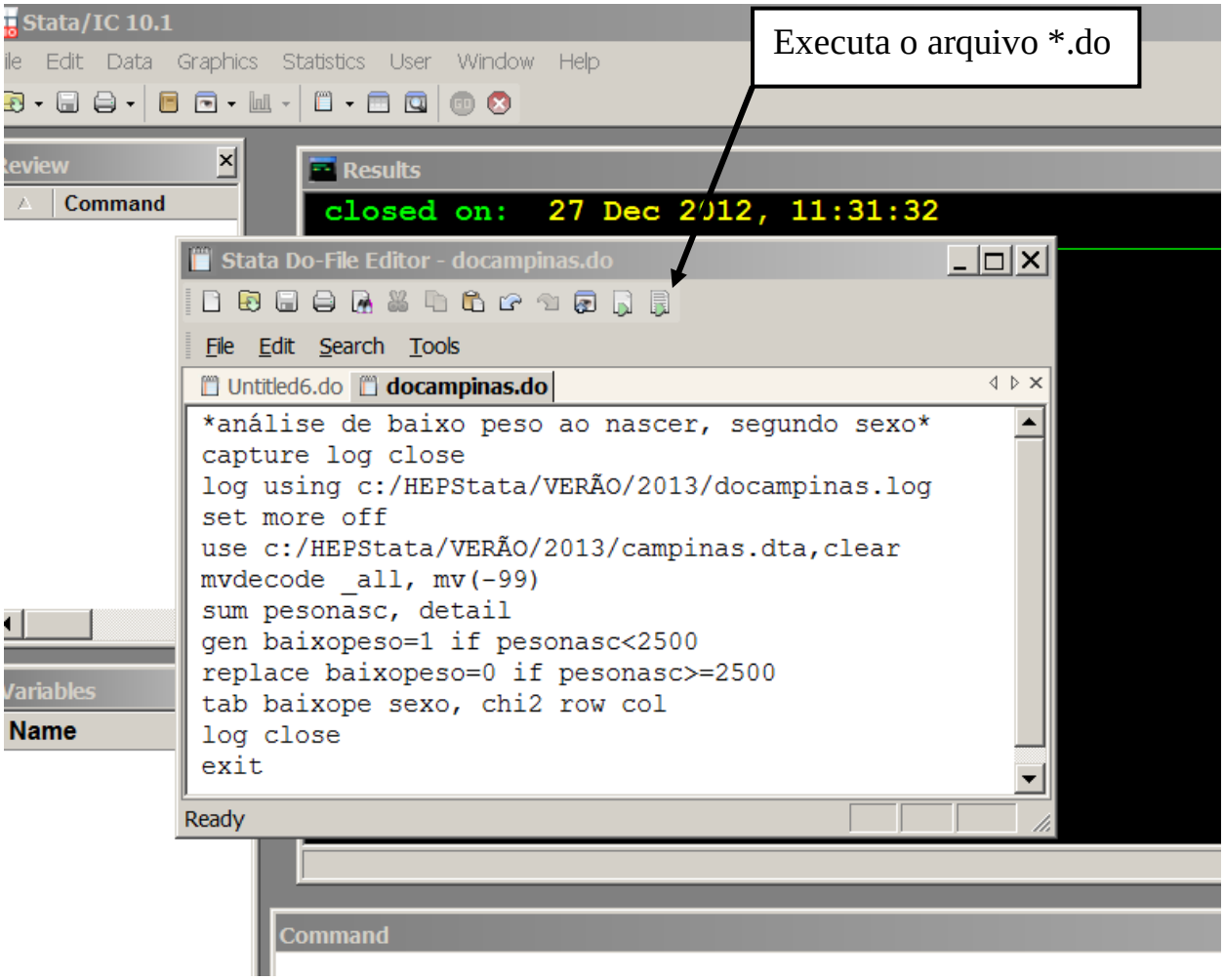
- **do <nome do arquivo de programação (.do)>**

Ou

Pressionar o botão ***do current file*** do *Do-editor*.

Esta mesma análise poderá ser feita para o banco de dados c:\...\botucatu.dta.





Aproveitando os comandos registrados na janela Review.

Botão da direita do mouse: Select all

Botão da direita do mouse: Sendo to Do-file editor

Uma janela de .do é aberta. Completar, salvar.

11. Arquivos não *.dta

11.1. A alternativa prática é usar o pacote comercial ***Transfer* (Stat/Transfer).**

11.2. Sugestão alternativa, criando planilha Excel.

11.2.1. Excel em inglês:

1. Construir planilha Excel, usando "." (ponto) como separador decimal.
2. Usar para registrar dado faltante algum código numérico com valores inteiros, talvez negativos (-9, -99 por exemplo).
3. Entrar datas no formato dd/mm/aaaa.
4. Salvar como "mesmonome.txt", separado por tabulação.

11.2.2. Excel em português:

1. Construir planilha Excel, usando ", " (vírgula) como separador decimal.
2. Usar para registrar dado faltante algum código numérico com valores inteiros, possivelmente negativos (-9, -99 por exemplo).
3. Entrar datas no formato dd/mm/aaaa.
4. Salvar como "mesmonome.txt", separado por tabulação.
5. Usar o bloco de notas e chamar "mesmonome.txt", fazer substituição de "," por "." e salvar.

Para leitura no Stata, usar o menu:

→ → →

File *Import* *ASCII data created by a spreadsheet*

→

→ → →

dataset filename, usando a caixa Browse botão *tab delimited*

caixa *replace* OK.

Sugestão para transferir do Stata para planilha Excel

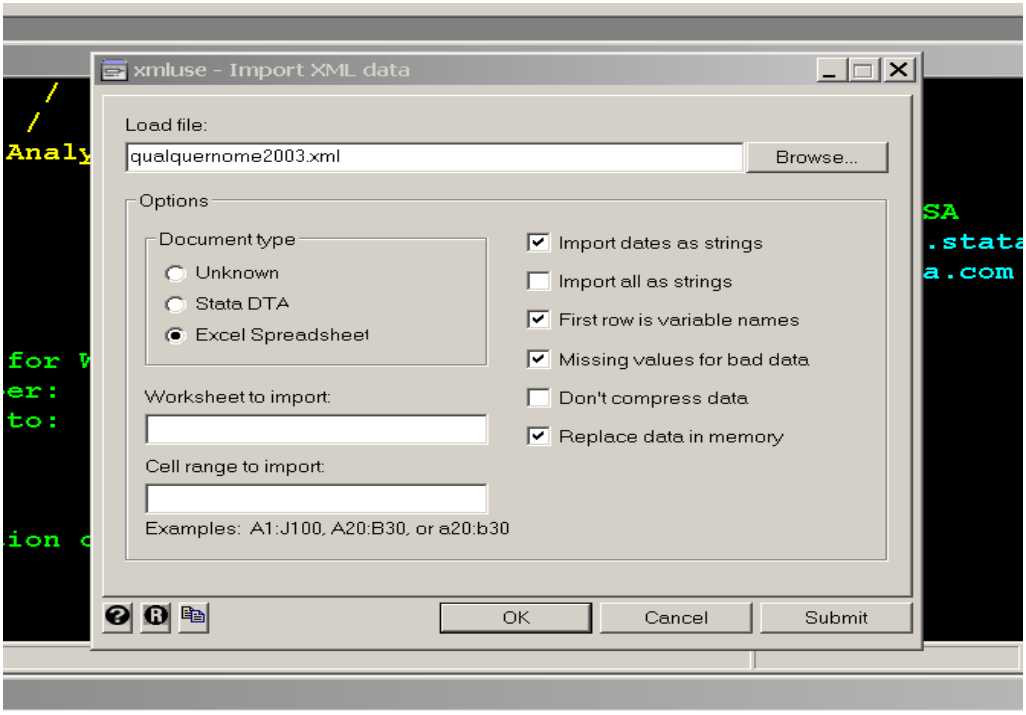
Usar o Transfer ou fazer o caminho contrário, com *Export*.

11.2.3. Excel em português, salvando como .xml:

1. Construir planilha Excel, usando “, “ (vírgula) como separador decimal.
2. Usar algum código numérico com valores inteiros, possivelmente negativos (-9, -99 por exemplo).
3. Entrar datas no formato dd/mm/aaaa.
4. Salvar como “qualquernome2003.xml”, em Tipo: Planilha XML 2003 (.xml).

Para leitura no Stata, usar o menu:

File → Import → XML data .



Obs: as vírgulas separadoras são transformadas automaticamente em pontos.

A seguir, no Stata:

- **compress**
- **save as “**qualquernome2003”
- **gen datanumérica=date(datastring,”DMY”)**

12. Exercícios

Exercício 1

1- iniciar o *Stata*

2- abrir um arquivo **exerc1.log** no sub-diretório c:\cursosta

3- abrir banco de dados existente em **C:\cursosta\fem.dta**

4- estudar as variáveis existentes utilizando o comando **describe**

5- alterar o banco de dados utilizando o Editor

paciente 2	age =43	anxiety =3
paciente 10	sleep=1	life= 1

quando terminar, salve as alterações (utilizando a opção **preserve**) e volte para a janela de comandos.

6- listar age

7- renomear o nome da variável *depress* para depressao

8- formatar a variável *weight* para 2 casas após a virgula

9- salvar o banco de dados como **c:\cursosta\femcorr.dta** (utilizando a opção **Save As** do menu)

10- fechar o arquivo de dados utilizando o comando **clear**

11- verificar se o arquivo **.log** continua aberto, utilizando o quarto ícone , visualizando-o.

12-fechar (suspender definitivamente) o arquivo **.log**

13-abrir arquivo de dados **c:\cursosta\breast.dta**

14- abrir arquivo **exerc1.log** como continuação (append) do arquivo

15- visualizar variáveis do banco utilizando o comando **describe**

16- listar os dados utilizando o comando **list**

17- fechar o arquivo de dados utilizando o comando **clear**

18- fechar o arquivo **exerc1.log**

19- abrir arquivo c:\cursosta\rim.dta

20- abrir um arquivo **.log (rim.log)**

21- substituir os valores codificados como -99 para valores faltantes (.)

22- recodificar a variável sexo, sendo 1=0 e 0=1

23- rotular as variáveis: **id "identificacao"**; **dias "tempo ate ocorrer o obito"**; **censura "condicao do paciente no fim do estudo"**; **tratam "tratamento"**; **doador "tipo de doador"**. Verifique se os labels foram criados corretamente através do comando describe

24- definir rótulos para as categorias das variáveis

variável	Codificação	
Sexo	0 – masculino	1 – feminino
Tratam	0 – sem imunossupressor	1– com imunossupressor
Doador	0– vivo	1 – cadáver

25- verificar os rótulos gerados utilizando o comando **tab** <nome da variável> (uma de cada vez)

26- pedir um resumo das variáveis utilizando o comando **summarize** ou **sum**

27- tabular os resultados de tratam, doador e sexo, com o comando **tab**.

28- gerar uma nova variável **idade_30** centrada na média utilizando o comando **gen idade_30 = idade – 30**

29- listar as variáveis **idade** e **idade_30**; verificar se a nova variável foi criada corretamente

30- gerar uma nova variável (**catidad)** que categorize a idade em:

Faixa etária	Código
10 |-- 21	1
21 |-- 31	2
31 |-- 41	3

≥ 41	4

Cuidado: valores *missing* serão categorizados na ultima categoria se não houver uma linha de comando específica para esta situação!!!

31-definir rótulos para as categorias de **catidad**

32- tabular a variável **catidad**

33- retirar a variável **idade_30**

34- fazer o teste de associação qui-quadrado entre as variáveis sexo e doador, com as porcentagens na linha.

35- fazer o teste de associação exato de Fisher entre as variáveis doador e tratamento, com as porcentagens na linha e coluna.

36- fazer o teste de diferenças de duas médias (t de "Student") para idade segundo tratamento

37- salvar o banco de dados incluindo a nova variável gerada utilizando o comando **save as rim2**, **replace**

38- fechar o arquivo **rim.log** e abrir no *Word*.

Gabarito – Exercício 1

1- pelo ícone ou **Iniciar**, **Programas**, **Stata**, **Intercooled Stata**

2- clicar no quarto ícone da barra de menu, mudar diretório para **c:\cursosta**, salvar com nome **exerc1.log**, fechar janela do arquivo **.log**

3- **use c:\cursosta\fem.dta** ou pelo menu, **File, Open** e seleciona-se o arquivo **fem.dta**, no diretório c:\cursosta

4- **describe** ou **desc**

5- utilizar o editor do Stata (8° ícone) para correção ou digitar **edit**. Após as mudanças salvar, clicando em **preserve**

6- **list**

7- **rename depress depressao**

8- **format weight %9.2f**

9- **File, Save As**. Salvar com o nome **femcorr.dta**

10- **clear**

11- clicar sobre o 4^0 ícone, escolher a 1ª. opção (**Bring log window to top**); rolar a tela do arquivo **.log**, fechar a janela do arquivo **.log**

12- clicar sobre o 4^0 ícone e selecionar a opção **Close log file**.

13- **use c:\cursosta\breast.dta** ou pelo menu, **File, Open** e seleciona-se o arquivo **breast.dta**, no diretório c:\cursosta

14- clicar no quarto ícone da barra de menu, mudar diretório para **c:\cursosta**, abrir o **exerc1.log**, fechar janela do arquivo **.log.** Escolher a opção append to existing file.

15- **describe** ou **desc**

16- **list**

17- clicar sobre o 4^0 ícone e selecionar a opção **Close log file**.

18- **clear**

19- **use c:\cursosta\rim.dta** ou pelo menu, **File, Open** e seleciona-se o arquivo **rim.dta**, no diretório c:\cursosta

20- clicar no quarto ícone da barra de menu, mudar diretório para **c:\cursosta**, salvar com nome **rim.log**, fechar janela do arquivo **.log**

21- **mvdecode _ all, mv(-99)**

22- **recode sexo 1=0 0=1**

23- **label variable id "identificação"**

label var dias "tempo até ocorrer o óbito"

label var censura "condição do paciente no fim do estudo"

label var tratam "tratamento"

label var doador "tipo de doador"

describe ou **desc**

24- **label define cen 0"censura" 1"falha"**

label val censura cen

label define s 0"masculino" 1"feminino"

label val sexo s

label define trat 0"sem imunossupressor" 1"com imunossupressor"

label val tratam trat

label define doa 0"vivo" 1"cadaver"

label val doador doa

25- **tab censura**

tab sexo

tab tratam

tab doador

ou **tab1 censura sexo tratam doador**

26- **sum** ou **summarize**

27- **gen idade_30=idade-30**

28- **list idade idade_30**

29- **gen catidad=1 if idade<21**

replace catidad=2 if idade>=21 & idade<31

replace catidad=3 if idade>=31 & idade<41

replace catidad=4 if idade>=41

replace catidad=. if idade==.

30- **label define catid 1 "menor que 20" 2 "20 a 30" 3 “30 a 40” 4 “maior que 40”**

label val catidad catid

31- **tab catidad**

32- **drop idade_30**

33-**tab sexo doador, chi2 row**

34- **tab tratam doador, exact row col**

35- **ttest idade, by(tratam)**

36- **Save as rim2, replace**

37-clicar sobre o 4^0 ícone, escolher a 1ª. opção (**Bring log window to top**); rolar a tela do arquivo **.log**, fechar a janela do arquivo **.log**

38- Minimizar ou fechar o Stata e abrir o *Word*. Abrir o arquivo rim.log da mesma maneira para abrir um arquivo documento.

Exercício 2

* Arquivo fem.dta

1. Faça o resumo da variável **weight** segundo nível de depressão (variável **depress**);

2. Faça a tabela que contém somente o peso médio e o desvio padrão da variável perda de peso (**weight**) para os níveis da variável **depress**;

3. Procure no **Help** a sintaxe do comando para realizar o *teste U de Mann-Witney*;

4. Compare as mudanças de peso segundo a variável **depress**, utilizando o *teste U de Mann-Witney*;

5. Faça um histograma da variável **age** e salve-o em um arquivo **doc**.

6. Faça um **boxplot** da variável **weight** segundo níveis da variável **depress.**

7. Transporte este gráfico para o *Word.*

Gabarito - exercício 2

1- sort depress

by depress: sum weight

2- table depress, contents(mean weight sd weight)

3- Help, Search. Digitar Mann-Whitney. Clicar na opção signrank (o teste de Mann-Whitney é feito pelo comando ranksum).

4- ranksum weight, by(life)

5- histogram age

Edit, Copy Graph. Abrir o Word, colar no documento e salvá-lo em um arquivo do Word.

6- graph box weight, over(depress) Edit, Copy Graph. Abrir o Word, colar no documento e salvá-lo em um arquivo do Word.

13. Miscelânea

Comandos que auxiliam o encontro de .ado escritos por usuários:

findit <algumpossívelcomando>

ssc install <*pkname*>

Comandos úteis (*.ado), escritos por usuários:

distinct

linkplot

renvars

spostado

stcascoh

www.ats.ucla.edu/stat/stata/ado/analysis

Exemplo 1

```
foreach var of varlist var1 var2 …{
recode `var' 0/1=1 2/3=2 4/5=3
}
```

Exemplo 2

```
foreach num of numlist 2 5 7/9 15{
display `num'
comando …  if …& criidadpes==`num'
comando.....
}
```

Exemplo 3

```
forvalues i= 9(1)12{
display `i'
commando … if…& criidadpes==`i'
}
```

Exemplo 4 (contribuição de Scott Merryman a partir do comando *linkplot* de Nick Cox)

usar o banco c:/data/tiago.dta

……..

preserve

tempvar last

bysort grupo id (age): gen byte `last’==_n==_N

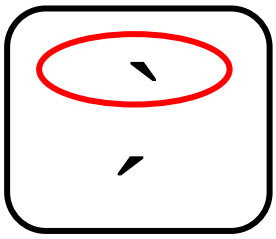
expand 2 if `last’

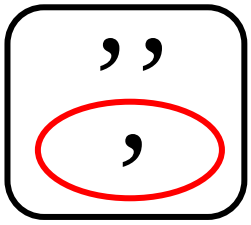
bysort grupo id (age): replace zpi=. if _n==_N

twoway scatter zpi age if grupo==0, cmissing(n) con(L) sort(grupo id age)||…

restore

…….





14. Bibliografia

Conrad S *Assignments in Applied Statistics*. Wiley, Chichester, 1989 (p.126).

Hamilton LC *Statistics with Stata: version 12*. Stata Press, 2013 (8th edit.)

Hand DJ et al. *A Handbook of Small Data Sets*. Chapmen e Hall, London, 1994.

Lea AJ New observations on distribution of neoplasms of female breast in certain European countries. *British Medical Journal*, 1, 488-490, 1965.

Mitchell MN *Stata Graphics – a visual guide.* Stata Press, 2008 (3th edit.:2012).

Rabe-Hesketh SR et Skrondal A *Multilevel and Longitudinal Modeling Using Stata*. Stata Press, 2008 (3th edit.: 2012).

Rabe-Hesketh SR et Everitt B A Handbook of Statistical Analysis Using Stata. Chapman & Hall, 2004 (4th edit.: 2007).

StataCorp. *Stata Statistical Software: releases 9/10.* Stata Corporation, 2004/2007.