

Microeconometria Utilizando Stata 11.0

Manual Stata 11.0

Esta apostila apresenta comandos básicos para manipulação de bases de dados com a utilização do aplicativo STATA 11.0 e introduz alguns conceitos básicos de estatística referentes aos comandos utilizados.

O leitor interessado em conhecer mais sobre este programa ou aprender teoria estatística mais detalhada deve procurar referências especializadas.

Introdução ao STATA 11.0

O STATA possui amplo potencial de utilização e trabalha com bases de dados que ficam armazenadas inteiramente na memória RAM do microcomputador. Por esta razão fornece processamentos de maneira muito rápida.

Em geral, os comandos do STATA tem a forma:

comando nomevar(s) **if....in...., options**

O STATA diferencia letras maiúsculas das minúsculas. Use sempre letras minúsculas quando digitar comandos, e recomendamos que você também use letras minúsculas para os nomes de suas variáveis. O STATA aceita abreviações para comandos e nomes de variáveis, desde que estas abreviações não sejam ambíguas.

Iniciando o STATA

O programa STATA, é iniciado clicando duas vezes no ícone localizado no desktop do Windows.

Janelas do STATA

Quatro janelas são apresentadas quando o STATA é iniciado. São elas:

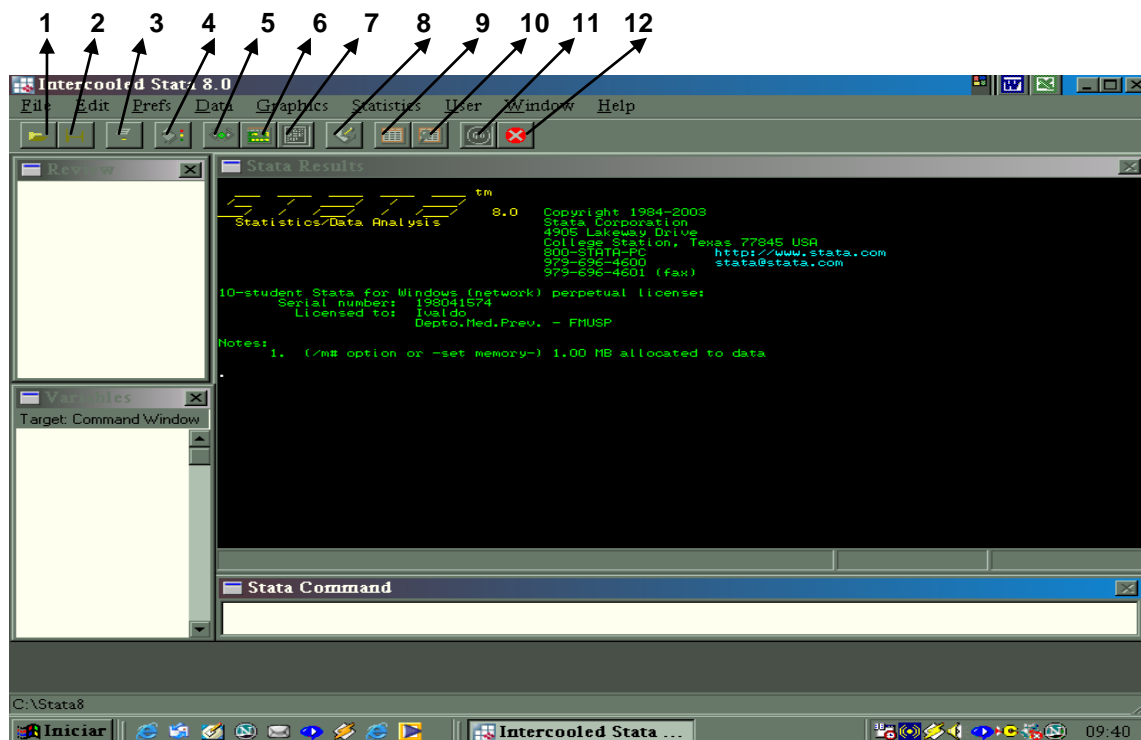
Review: janela onde são armazenados os comandos

Variables: janela que apresenta a lista das variáveis do banco de dados ativo

Stata Results: janela que mostra os resultados

Stata Command: janela onde os comandos do STATA devem ser digitados

Clicando com o botão direito do "mouse", na janela **Review**, ativa-se a opção para salvar os comandos.



O *menu* está disponível na primeira linha e possui os recursos:

File Edit Prefs Data Graphics Statistics Window e Help

Por exemplo, o *menu* "HELP ⇒ SEARCH" é utilizado para procurar ajuda sobre comandos do STATA.

Na segunda linha encontra-se a *Barra de Ferramentas* com os ícones:

- (1) **Open (use)**: Carrega ou abre um banco de dados no formato do STATA (dta).
- (2) **Save**: Salva um arquivo no formato do STATA (dta).
- (3) **Print Results**: Imprime a janela de resultados.
- (4) **Begin Log**: Carrega, abre ou cria um arquivo do tipo ".log" ou ".smcl".
- (5) **Start Viewer**: Exibe a tela de ajuda (Help) em primeiro plano.
- (6) **Bring Results Window to Front**: Exibe a tela dos resultados em primeiro plano.
- (7) **Bring Graph Window to Front**: Exibe a tela com o gráfico em primeiro plano.
- (8) **Do-file Editor**: Edita um arquivo de comandos (arquivo tipo ".do").
- (9) **Data Editor**: Edita o arquivo de dados que está sendo utilizado.
- (10) **Data Browser**: Visualiza o arquivo de dados que está sendo utilizado.
- (11) **Clear**: prossegue a execução do comando.
- (12) **Break**: Interrompe a execução de uma tarefa ou comando.

Tipos de arquivos do STATA

.ado	arquivos programa "do-files"
.dct	arquivos ASCII , arquivo dicionário
.do	do-file
.dta	arquivos de dados formato do STATA
.gph	arquivos gráficos
.log ou .smcl	arquivos textos com os resultados
.out	arquivos para impressão
.raw	arquivos ASCII arquivos de dados
.sum	arquivos controle de rede

Preparando o STATA para análise

Em primeiro lugar, você deve escolher o banco de dados que irá trabalhar e abrir/carrega-lo no STATA. Note que o STATA só abre bancos de dados no formato “.dta”, ou o banco pode ser copiado diretamente do formato XLSX para o STATA. Por isso, caso seu banco de dados não esteja neste formato, antes de iniciar o STATA você deve convertê-lo utilizando o programa STAT/TRANSFER que é um programa muito útil e fácil de ser usado. Na versão do Stata 11.0 você tem a opção de copiar o banco direto.

Para ilustrar, vamos trabalhar com o arquivo **xxxx**. Abra o banco de dados clicando no ícone **(1)Open** e, então, selecione o caminho (pasta) onde está o arquivo **xxx**. Note o que mudou nas janelas do STATA!!

Agora, vamos abrir também um arquivo do tipo “**log**” onde ficarão armazenados todos os resultados gerados a partir da tela de comandos. Isto pode ser feito clicando o ícone **(4)Begin Log** e, então, selecionando o tipo de arquivo=**log**, o caminho e o nome para o arquivo.

Comandos básicos do STATA :

describe	descreve o arquivo de dados em uso
display	calculadora de mão
drop	elimina variáveis ou observações

edit	edita e lista dados
generate	cria ou muda conteúdos de variáveis
graph	cria gráficos
list	lista os valores das variáveis por registro
memory	muda o tamanho da memória a ser utilizada
obs	aumenta o número de observações no banco de dados
recode	recodificar, agrupar códigos
sort	ordena os dados
summarize	calcula medidas de tendência central
tabulate	produz tabelas simples e cruzadas

Utilize o *help* do STATA para obter mais informações sobre estes e outros comandos.

Salvando os comandos

Todos os comandos digitados na janela **STATA Command** são enviados para a janela **Review**. Estes comandos podem ser guardados em um arquivo especial (arquivos tipo "**do**") para, posteriormente, ser editado e utilizado em uma nova análise.

Para criar um arquivo do tipo "**do**" utilize o botão direito do "mouse" na janela **Review**.

Análise descritiva

Após a coleta de dados e a digitação dos mesmos em um banco de dados apropriado, o próximo passo é a análise descritiva. Esta etapa é fundamental, pois uma análise descritiva detalhada fornece ao pesquisador toda a informação contida no conjunto de dados. Neste enfoque, procura-se obter a maior quantidade possível de informação, buscando responder às questões que estão sendo pesquisadas.

As variáveis podem ser classificadas em contínuas ou categóricas. Por **variável contínua** (ou quantitativa) entende-se as variáveis que podem assumir todos os

valores possíveis dentro de um limite especificado. **Variável categórica** (ou qualitativa) é aquela que pode ser classificada em categorias separadas e que não assumem valores intermediários, como por exemplo, sexo e estado civil.

Em geral, uma análise descritiva dos dados é feita com base em medidas de posição e variabilidade. Para variáveis contínuas, as medidas comumente utilizadas são as medidas de tendência central, enquanto as variáveis categóricas são sumarizadas por meio de medidas de frequência.

Medidas de tendência central:

média aritmética: é a soma de todas as observações dividida pelo número de observações.

mediana: valor central de uma distribuição. Para se obter a mediana, ordena-se as observações em ordem crescente. Se o número de observações for par, a mediana será a média aritmética dos dois valores centrais ($n/2$ e $[(n/2)+1]$), onde n é o número de observações total da amostra. Se o número de observações for ímpar, a mediana será o valor na posição $(n + 1)/2$.

moda: é o valor com a maior frequência entre todas as observações.

frequência: é o número de vezes em que um valor ocorre.

A seguir são apresentados alguns comandos básicos para fazer uma análise descritiva dos dados:

Aplicação prática-1 - Digitando os comandos na janela **Command**

Digite **describe** ou **desc** e pressione ENTER, deve aparecer na janela **Stata Results** o seguinte resultado:

```
Contains data from C:\Motocobr.dta
  obs:                800
  vars:                 18                22 Aug 2000 15:44
  size:               35,200 (96.3% of memory free)
```

```
-----
```

1. id	long	%12.0g		id
2. idade	byte	%8.0g		idade
3. pausas	byte	%8.0g		numero de pausas dia
4. escola	long	%19.0g	escola	escola
5. nasc	byte	%8.0g	nasc	procedencia
6. tsp	int	%11.0g	tsp	tempo em SP
7. emp	int	%8.0g	emp	tipo de empresa
8. fun	int	%9.0g	fun	funcao
9. esc	int	%13.0g	esc	escala da linha do onibus
10. fol	int	%8.0g	fol	escala de folga
11. jorn	int	%11.0g	jorn	jornada de trabalho diaria
12. temp	int	%9.0g	temp	tempo de trabalho na empresa
13. trans	long	%12.0g	trans	transito diario
14. banco	long	%12.0g	banco	possibilidade ajuste do assento
15. fal	int	%8.0g	fal	falta ao trabalho no ultimo mes
16. sono	int	%10.0g	sono	sono diario
17. tmc	int	%8.0g	srq	transtorno mental comum
18. sal	byte	%8.0g	sal	salario mensal

```
-----
```

Sorted by:

Digite **list in 1** e pressione ENTER

Observation 1

id	27	idade	35	pausas	2
escola	primario com	nasc	nordeste	tsp	11-20 anos
emp	privada	fun	motorista	esc	linha altern
fol	muda	jorn	> 9	temp	< 4 anos
trans	intenso	banco	sim	fal	nao
sono	>= 6 horas	tmc	nao	sal	> 6 sm

Para mudar o nome de uma variável, como por exemplo, id para identif, digite

rename id identif

e pressione ENTER

Para observar a mudança. Digite **desc**

Os comandos **tabulate** , **tab** ou **tab1** produzem tabelas simples ou cruzadas.

tab escola

escola	Freq.	Percent	Cum.
-----+-----			
ginasio completo	84	10.50	10.50
primario completo	554	69.25	79.75
primario incompleto	162	20.25	100.00
-----+-----			
Total	800	100.00	

```
tab escola, nolabel
```

escola	Freq.	Percent	Cum.
0	84	10.50	10.50
1	554	69.25	79.75
2	162	20.25	100.00
Total	800	100.00	

Agora digite: **tab1 escola fun emp**

Deve aparecer na tela os seguintes resultados:

-> tabulation of escola

escola	Freq.	Percent	Cum.
ginasio completo	84	10.50	10.50
primario completo	554	69.25	79.75
primario incompleto	162	20.25	100.00
Total	800	100.00	

-> tabulation of fun

funcao	Freq.	Percent	Cum.
motorista	423	52.88	52.88
cobrador	377	47.12	100.00
Total	800	100.00	

-> tabulation of emp

tipo de empresa	Freq.	Percent	Cum.
publica	286	35.75	35.75
privada	514	64.25	100.00
Total	800	100.00	

Para criar a variável **nasc2**, recodificar e inserir um rótulo (*label*), utilize os comandos:

```
tab nasc (tabela de freqüência)
gen nasc2=nasc (criar variável nasc2)
recode nasc2 0=0 1=0 2=1 3=1 (recodifica variável nasc2)
```

```
label var nasc2 "Grupos de Procedência" (insere var-label)
```

```
tab nasc2
label define cod_proc 0 "sudeste" 1 "outros" (insere value-label)
```

```
label val nasc2 cod_proc
tab nasc2
```


Seus resultados devem ter sido:

tab nasc

procedencia	Freq.	Percent	Cum.
SP	281	35.12	35.12
RJ/MG/ES	135	16.88	52.00
outros	48	6.00	58.00
nordeste	336	42.00	100.00
Total	800	100.00	

tab nasc2

Grupos de Procedência	Freq.	Percent	Cum.
0	416	52.00	52.00
1	384	48.00	100.00
Total	800	100.00	

tab nasc2

Grupos de Procedência	Freq.	Percent	Cum.
sudeste	416	52.00	52.00
outros	384	48.00	100.00
Total	800	100.00	

O comando **summarize** ou **sum** é utilizado para calcular média, desvio padrão, mínimo, máximo, etc.

summarize idade

Variable	Obs	Mean	Std. Dev.	Min	Max
idade	800	37.69	10.52532	17	67

sum idade, detail

idade					
Percentiles		Smallest			
1%	21	17			
5%	22	19			
10%	24	19	Obs		800
25%	30	19	Sum of Wgt.		800
50%	37		Mean		37.69
		Largest	Std. Dev.		10.52532
75%	45	65			
90%	53	66	Variance		110.7824
95%	58	66	Skewness		.440607
99%	63	67	Kurtosis		2.555018

Aplicação prática-2 - Utilizando os menus: **Data** e **Statistics**

2.1 - Para descrever o arquivo e suas variáveis, clique no menu **Data**, opção: Data describe e explore todas as opções.

2.2 – Para editar o banco de dados, clique no menu **Data**, opção: Data editor .

2.3 – Para produzir tabelas simples, clique no menu **Statistics**, opção: Summaries, tables & tests → Tables .

Explore as opções do Summaries.

2.4 Criar a variável nasc3 a partir da var´ NASC. Clique no menu **Data**, opção: Create or change variables → Create new variable

Acrescente label para a var´ nasc3 e label para os valores da var´ nasc3

(menu Data, opção: Label & notes)

2.5 Calcular a média, mediana, desvio padrão, ... Clique no menu **Data**, opção Describe data → Summary statistics ou menu **Statistics**, opção: Summaries, tables & tests → Summary statistics.

Aplicação prática-3

3.1 – Ler/abrir o arquivo : *motocobr.dta* Clique no menu **File**, opção *open*
Abrir arquivo(**log**) para armazenar os resultados: Clique no botão: *Begin log*
Digite *motcob* em nome do arquivo e clique e no botão *SALVAR*.

3.2 – Produzir tabela de freqüência simples para as variáveis *IDADE* e *FUN*.

Comandos : **tab1 idade fun**
tab1 idade fun , nolabel

3.3 – Criar nova var´ *IDADER*, *idade* recodificada(agrupada) nas faixas:
ate 30 ; 31 a 40 ; 41 a 50 ; 51 e mais

Comandos: **generate idader= idade**
recode idader 17/30=1 31/40=2 41/50=3 51/67=4 ou
recode idader min/30=1 31/40=2 41/50=3 51/max=4

outra maneira : **recode idade 17/30=1 31/40=2 41/50=3 51/67=4, gen(idade1)**
ou **egen idade2 = cut(idade), at (17, 31, 41, 51, 68)**

3.4 - Inserir *labels* para a variável *IDADER*

Comandos: **label var idader "idade agrupada"**
label define cod_idade 1 " ate 30" 2 "31 - 40" 3 "41 - 50" 4 "51 e mais"
label val idader cod_idade

3.5 - Produzir tabelas:

Comandos: **tab idader**
tab idader fun
tab idader fun , row col cel chi

3.6 - Salvando os arquivos:

(1) arquivo de dados(**dta**) : menu **File**, opção Save as ... e digite *motcob* em nome do arquivo.

(2) arquivo de resultados(**log**) : clique no botão : *Close/suspend Log* e escolha a opção *close log file* e clique no botão *OK* para salvar.

Abra o arquivo motcob.log no Word e observe os resultados.

(3) arquivo de comandos(**do**) : clique na da janela **Review** com o botão direito do mouse , e escolha a opção *Save Review Contents ...* e digite *motcob* em nome do arquivo.

Visualizando o arquivo do : clique no botão *Do-file Editor* para abrir o arquivo *motcob.do*.

Correndo(Run) o arquivo do: digite o comando abaixo na janela *Command* para correr(processar) os comandos do arquivo **do** :

do c:\stata básico\motcob.do

Gráficos

O comando `graph` do STATA possui várias opções. Em geral, gráficos de barra são usados para mostrar a distribuição de variáveis categóricas, enquanto histogramas e box-plots são usados para mostrar a distribuição das variáveis quantitativas.

Para obter um gráfico de barras da variável `escola`, utilize:

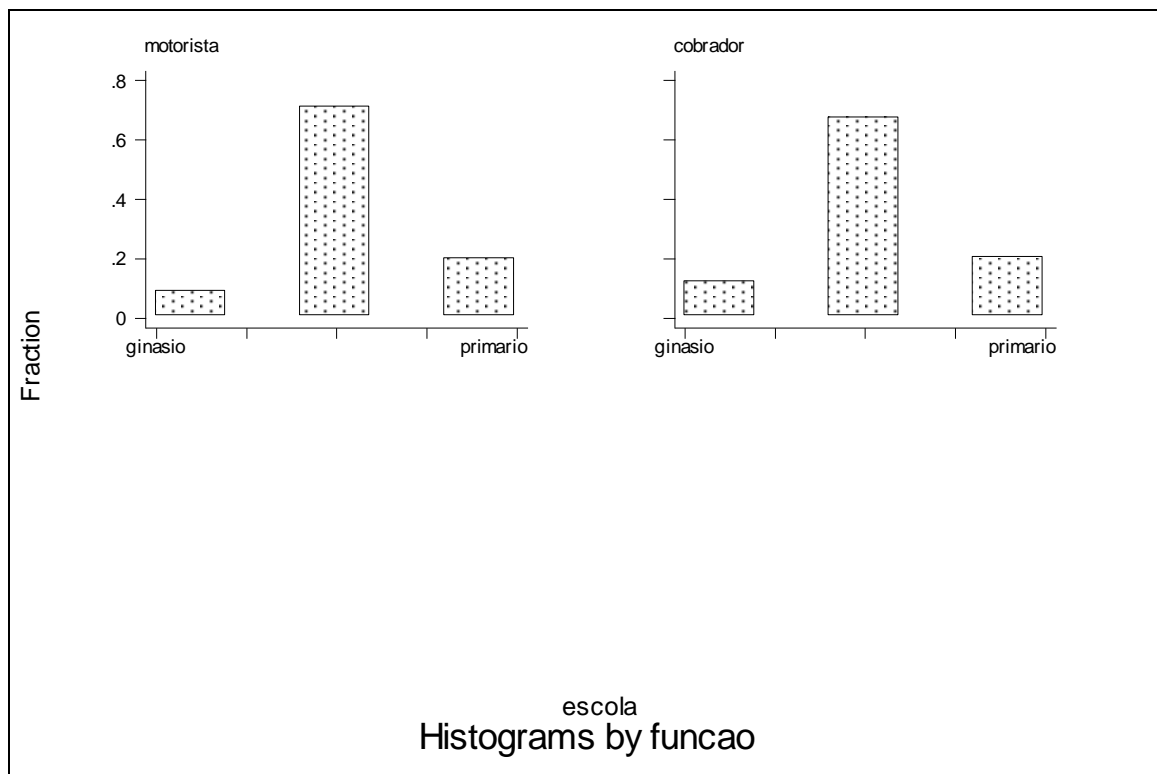
```
graph7 escola          ou          gr7 escola
```

Você pode usar também a opção **ylabel**

```
graph7 escola, ylabel
```

Ou ainda, obter o mesmo gráfico para cada estrato da variável função

```
sort fun  
graph7 escola, ylabel by(fun)
```



A cada novo gráfico que o Stata gerar, o anterior será "perdido", por isso, às vezes é desejável salvar um gráfico antes de gerar outro. Um gráfico pode ser salvo de duas maneiras diferentes: a primeira é copiar cada gráfico e colar em um outro arquivo "fora"

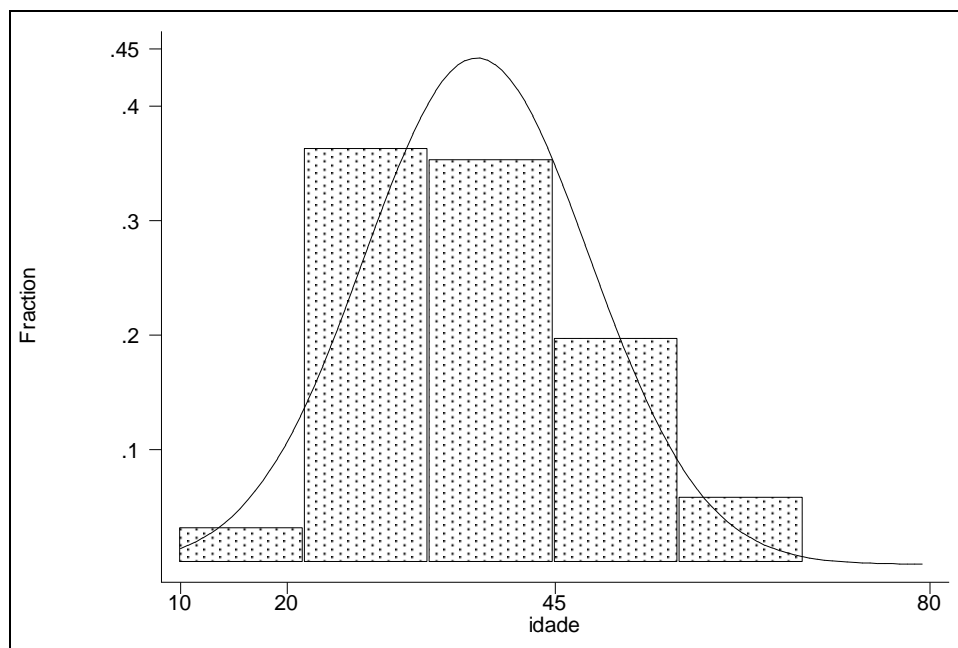
do STATA, por exemplo, um arquivo do Word. Uma outra maneira é salvar a janela com o gráfico como uma figura, utilizando o *menu*: File → Save graph .

Para obter um histograma da variável IDADE, digite:

```
graph7 idade, hist ou histogram idade ou hist idade
```

Para melhorar a apresentação visual do histograma, utilize a opção *xlabel* e *ylabel*. O número de retângulos do histograma pode ser modificado pela opção *bin(x)*. Para sobrepor ao seu histograma uma curva normal com média e desvio padrão, adicione a opção *normal*.

```
gr7 idade, hist xlabel ylabel bin(10) normal freq
```

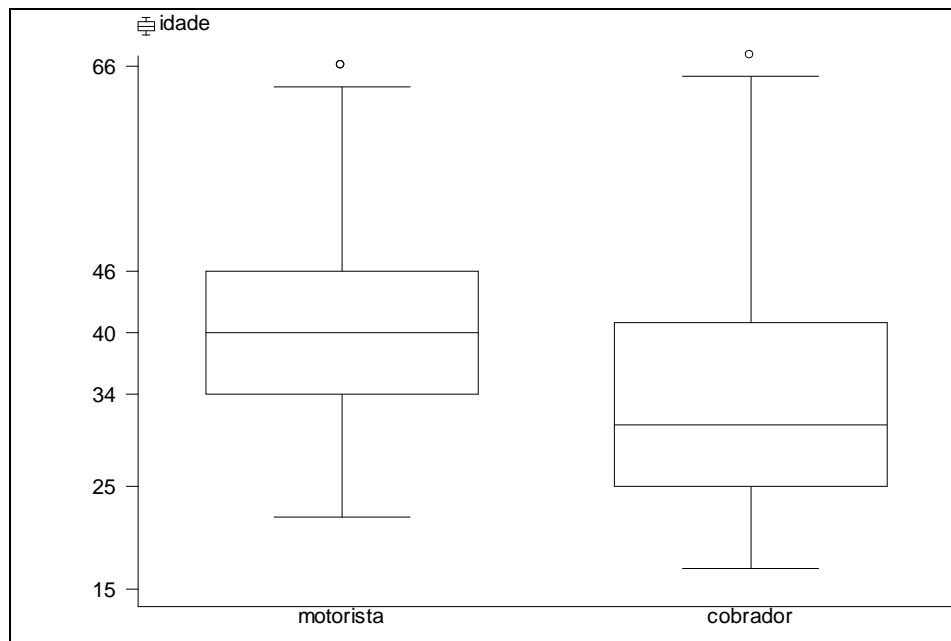


Um gráfico do tipo boxplot para a variável IDADE pode ser obtido com o comando

```
gr7 idade, box
```

Você pode ainda, construir este mesmo gráfico segundo FUN

```
sort fun
gr7 idade, box by(fun)
gr7 idade, box by(fun) ylabel(15,25,34,40,46,66)
```



Criando um arquivo “do” para obter gráfico de barras. Clique no botão *Do-file Editor*, digite as linhas abaixo e salve com o nome graf1.do .

```
clear
input  cobradores  motoristas  str12  esc
45.24  54.76  "g.c."
54.15  45.85  "p.c."
52.47  47.53  "p.i."
end
sort  esc
gr bar  cobradores  moto, over(esc)
```

Para correr(run) do arquivo **do** , digite o comando abaixo na janela *Command*

do graf1

Testes de hipóteses

Testes de hipóteses consistem em testar a significância estatística e quantificar o grau em que a variabilidade da amostra pode ser responsável pelos resultados observados no estudo. Para isto, define-se uma hipótese nula (H_0) e uma hipótese alternativa (H_a), que podem representar, por exemplo:

H_0 : não existe diferença entre exposição e doença

H_a : existe diferença entre exposição e doença.

Manipulação de variáveis categóricas

A seguir são ilustradas algumas maneiras de relacionar duas ou mais variáveis categóricas.

Relacionando duas variáveis categóricas

Suponha que você queira investigar se os trabalhadores que têm mais transtorno mental comum (TMC) faltam mais ao trabalho, ou seja, se existe uma associação entre TMC e a falta ao trabalho. Para isto, você pode construir uma tabela 2X2 usando o comando `tabulate` ou, de forma abreviada, `tab`

```
tab tmc fal
```

Somente esta tabulação não mostra com clareza se as duas variáveis analisadas estão associadas.

O que podemos fazer?

Uma opção simples é analisar as porcentagens destas variáveis em relação aos totais observados. Os subcomandos `row`, `col` e `cel` fornecem, respectivamente, as porcentagens das linhas, colunas e do total:

```
tab tmc fal, col
tab tmc fal, row
tab tmc fal, cel
```

Uma outra maneira de avaliar a associação é utilizando um teste de associação.

Avaliando a associação de duas variáveis com o teste Qui-quadrado de Pearson

Ainda com o objetivo de estudar a associação entre *função do empregado* e *presença de falta no último mês*, vamos usar o teste Qui-quadrado de Pearson para testar a significância da associação. Para isto, utilize a opção **chi**.

```
tab tmc fal, row chi
```

transtorno mental comum	falta ao trabalho no ultimo mes		Total
	nao	sim	
nao	485	160	645
	75.19	24.81	100.00
sim	100	55	155
	64.52	35.48	100.00
Total	585	215	800
	73.13	26.88	100.00

```
Pearson chi2(1) = 7.2500 Pr = 0.007
```

Considerações a respeito da validade do teste Qui-quadrado de Pearson

O teste Qui-quadrado de Pearson segue, aproximadamente, uma distribuição chamada Qui-quadrado (χ^2). Para amostras grandes esta suposição é razoável. No entanto, as seguintes regras podem ser usadas para garantir a validade do uso do teste:

- para tabelas 2 x 2, o teste χ^2 pode ser usado :
 - se o tamanho total da amostra (N) é maior do que 40,
 - se N está entre 20 e 40 e o menor valor esperado é maior ou igual a 5
- para tabelas de dimensões maiores :
 - o teste χ^2 é válido se não mais do que 20% dos valores esperados forem menores do que 5 e nenhum for menor do que 1.

Caso o teste χ^2 não seja adequado, uma opção é utilizar o teste exato de Fisher obtido com o subcomando **exact**.

```
tab tmc fal, row exact
```

transtorno mental comum	falta ao trabalho no ultimo mes		Total
	nao	sim	
nao	485	160	645
	75.19	24.81	100.00
sim	100	55	155
	64.52	35.48	100.00
Total	585	215	800
	73.13	26.88	100.00

```

Fisher's exact = 0.009
1-sided Fisher's exact = 0.005

```

O que você conclui da associação acima?

Relacionando três variáveis categóricas

Utilize os comandos **if**, **by** ou **tabulate**, como mostrado a seguir:

```
tab tmc fal if fun==1, row chi
```

transtorno mental comum	falta ao trabalho no ultimo mes		Total
	nao	sim	
nao	203	76	279
	72.76	27.24	100.00
sim	62	36	98
	63.27	36.73	100.00
Total	265	112	377
	70.29	29.71	100.00

```
Pearson chi2(1) = 3.1308 Pr = 0.077
```

```
tab tmc fal if fun==0, row chi
```

```
. tab tmc fal if fun==0, row chi
```

transtorno mental comum	falta ao trabalho no ultimo mes		Total
	nao	sim	
nao	282	84	366
	77.05	22.95	100.00
sim	38	19	57
	66.67	33.33	100.00
Total	320	103	423
	75.65	24.35	100.00

```
Pearson chi2(1) = 2.8861 Pr = 0.089
```

E agora? Qual a sua conclusão sobre a associação entre TMC e falta ao trabalho?

Manipulação de variáveis contínuas

Construção de intervalos de confiança para a média

A média é uma medida pontual e não fornece nenhuma informação a respeito da variabilidade dos dados. Este procedimento não permite julgar qual a possível magnitude do erro que estamos cometendo. Daí surge a idéia de construir o intervalo de confiança, que é definido como o intervalo dentro do qual se encontra a verdadeira magnitude do efeito com um certo grau de certeza.

O comando abaixo ilustra a construção do intervalo de confiança (IC) para a média da variável **idade**.

```
ci idade
```

Variable	Obs	Mean	Std. Err.	[95% Conf. Interval]	
idade	800	37.69	.3721263	36.95954	38.42046

Com base na amostra deste estudo, podemos dizer, com 95% de confiança, que o verdadeiro valor para a idade média dos motoristas e cobradores está entre 37,0 e 38,4 anos.

Note que, quando não especificamos um determinado nível de confiança, o programa assume $\gamma = 95\%$ para o cálculo do intervalo. No entanto, é possível mudar este valor usando a opção `level`.

No exemplo abaixo, o IC foi construído com confiança de 90%.

```
ci idade, level(90)
```

Variable	Obs	Mean	Std. Err.	[90% Conf. Interval]	
idade	800	37.69	.3721263	37.0772	38.3028

O IC também pode ser utilizado para testar se a média de interesse é estatisticamente igual, com um certo coeficiente de confiança, a um determinado valor de interesse.

De maneira análoga, podemos fazer um teste de hipótese para avaliar a mesma questão: “Será que a idade média dos motoristas e cobradores é estatisticamente diferente de 35 anos?”

Para isto, podemos usar o teste t de Student :

```
ttest idade = 35
```

One-sample t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
idade	800	37.69	.3721263	10.52532	36.95954	38.42046

Degrees of freedom: 799

Ho: mean(idade) = 35

Ha: mean < 35
t = 7.2287
P < t = 1.0000

Ha: mean ~= 35
t = 7.2287
P > |t| = 0.0000

Ha: mean > 35
t = 7.2287
P > t = 0.0000

Comparação entre médias de duas amostras independentes

Suponha agora que você queira avaliar se a idade média difere segundo a função do trabalhador. Neste caso, utiliza a opção `by(fun)`:

ttest idade, by(fun)

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
motorist	423	40.74468	.4227253	8.694175	39.91377	41.57559
cobrador	377	34.2626	.5833693	11.32698	33.11552	35.40967
combined	800	37.69	.3721263	10.52532	36.95954	38.42046
diff		6.482081	.7097834		5.088818	7.875344

Degrees of freedom: 798

Ho: mean(motorist) - mean(cobrador) = diff = 0

Ha: diff < 0	Ha: diff ~= 0	Ha: diff > 0
t = 9.1325	t = 9.1325	t = 9.1325
P < t = 1.0000	P > t = 0.0000	P > t = 0.0000

Considerações a respeito da validade do teste t de Student

O teste t assume que a distribuição da variável resposta é aproximadamente normal e o desvio padrão é o mesmo em cada grupo a ser comparado.

Então, no caso acima, estamos assumindo que o desvio padrão da variável IDADE (variável resposta) é o mesmo para motoristas e cobradores. Esta suposição precisa ser verificada, o que pode ser feito com o comando:

sdtest idade, by(fun)

Variance ratio test

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
motorist	423	40.74468	.4227253	8.694175	39.91377	41.57559
cobrador	377	34.2626	.5833693	11.32698	33.11552	35.40967
combined	800	37.69	.3721263	10.52532	36.95954	38.42046

Ho: sd(motorist) = sd(cobrador)

F(422,376) observed	= F_obs	=	0.589
F(422,376) lower tail	= F_L	= F_obs	= 0.589
F(422,376) upper tail	= F_U	= 1/F_obs	= 1.697

Ha: sd(1) < sd(2)	Ha: sd(1) ~= sd(2)	Ha: sd(1) > sd(2)
P < F_obs = 0.0000	P < F_L + P > F_U = 0.0000	P > F_obs = 1.0000

Quando o teste acima (teste de homocedasticidade) indicar que as variâncias não são iguais nos dois grupos, devemos usar um teste que considere esta desigualdade. Isto pode ser feito com o uso da opção **unequal**:

```
ttest idade, by(fun) unequal
```

Two-sample t test with unequal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
motorist	423	40.74468	.4227253	8.694175	39.91377	41.57559
cobrador	377	34.2626	.5833693	11.32698	33.11552	35.40967
combined	800	37.69	.3721263	10.52532	36.95954	38.42046
diff		6.482081	.7204279		5.06763	7.896533

Satterthwaite's degrees of freedom: 702.063

Ho: mean(motorist) - mean(cobrador) = diff = 0

Ha: diff < 0	Ha: diff ~= 0	Ha: diff > 0
t = 8.9975	t = 8.9975	t = 8.9975
P < t = 1.0000	P > t = 0.0000	P > t = 0.0000

No exemplo acima, qual dos testes você julga ser o mais correto? Por quê?
As idades médias dos motoristas e dos cobrados é estatisticamente igual?

Comparação entre médias de duas amostras dependentes

Quando as amostras não são independentes dizemos que as observações são correlacionadas e neste caso, o teste *t-pareado* é mais indicado pois leva em conta a correlação existente entre as observações.

Um exemplo de amostras dependentes é o estudo onde dois observadores diferentes fizeram medições da prega cutânea de 15 indivíduos distintos. As medidas são observadas no mesmo indivíduo, portanto, dizemos que as amostras dos 2 observadores são dependentes.

O banco de dados do estudo descrito anteriormente chama-se **Prega.dta**. Neste arquivo, os valores foram cadastrados de modo que cada indivíduo tem seus dados representados em uma linha diferente. As variáveis são descritas a seguir:

id = identificação do indivíduo

observA = medida da prega cutânea segundo o observador A

observB = medida da prega cutânea segundo o observador B

Para realizar o teste t-pareado basta digitar

```
ttest observa=observb
```

Paired t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
observa	15	23.84667	2.041145	7.905321	19.46885	28.22449
observb	15	21.56667	1.842221	7.134891	17.6155	25.51784
diff	15	2.28	.5819672	2.253949	1.031805	3.528196

Ho: mean(observa - observb) = mean(diff) = 0

Ha: mean(diff) < 0

t = 3.9177

P < t = 0.9992

Ha: mean(diff) ~= 0

t = 3.9177

P > |t| = 0.0015

Ha: mean(diff) > 0

t = 3.9177

P > t = 0.0008

Observando o resultado acima, o que você conclui?

Conversão de banco de dados - programa Stata/Transfer

Como foi comentado anteriormente, o STATA trabalha apenas com bancos de dados no formato "*dta*". O banco de dados que iremos utilizar agora (***Plasma.xls***) está no formato EXCEL e, portanto, deve ser convertido para o formato de um banco de dados do STATA. A conversão deve ser feita por meio do STATA/TRANSFER, que é um programa de conversão muito útil e fácil de ser usado. Antes de inicializar o STATA, utilize-o para converter o arquivo *Plasma.xls* em *Plasma.dta*.

Clique duas vezes no ícone **Stat Transfer** na área de trabalho

Na opção *transfer*, há as seguintes alternativas:

Input file type: das várias opções, escolha *Excel*

File specification: clique em *Browse* para achar o seu arquivo.

Note que, em seguida, o número de variáveis que foram automaticamente selecionadas é exibido.

Output file type: das várias opções, escolha *STATA*

File specification: automaticamente, o programa nomeia o arquivo com o mesmo nome do banco de dados original, porém com a extensão *dta*.

Clique em **Transfer**. Quando o programa terminar clique em EXIT.

Pronto! Seu banco pode ser utilizado no programa STATA!

Relação entre duas variáveis contínuas

1. Correlação linear de Pearson

Em muitas situações, é de interesse quantificar a força da relação linear entre duas variáveis contínuas, sem designar uma como resposta e outra como explicativa.

O grau desta associação pode ser medido com o uso do coeficiente de correlação linear de Pearson (r), que leva este nome pois foi descrito por Pearson. A correlação entre duas variáveis é positiva se valores mais altos de uma variável estão associados a valores mais altos da outra, e é negativa se os valores de uma variável crescem enquanto os da outra diminuem. O coeficiente de correlação próximo do zero significa que não existe uma relação linear entre as duas variáveis.

O coeficiente de correlação varia de -1 a $+1$, sendo:

+1: associação positiva perfeita

0: ausência de associação

-1: associação negativa perfeita

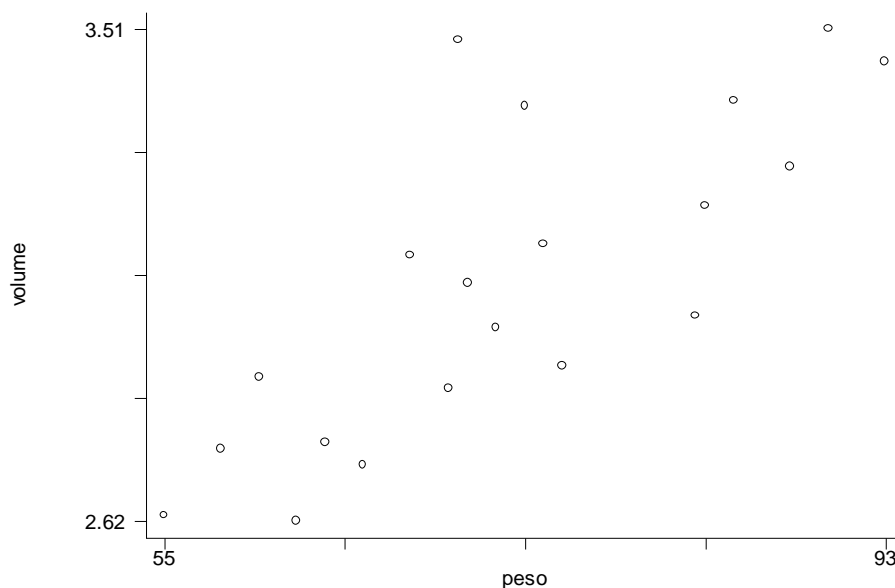
Aplicação prática:

Utilizando o banco de dados **plasma.dta** vamos verificar se existe uma relação linear entre as variáveis volume plasmático e peso.

A melhor forma de iniciar o estudo da possível relação entre estas duas variáveis contínuas é construir um gráfico de dispersão, utilizando os comandos:

```
graph7 volume peso
```

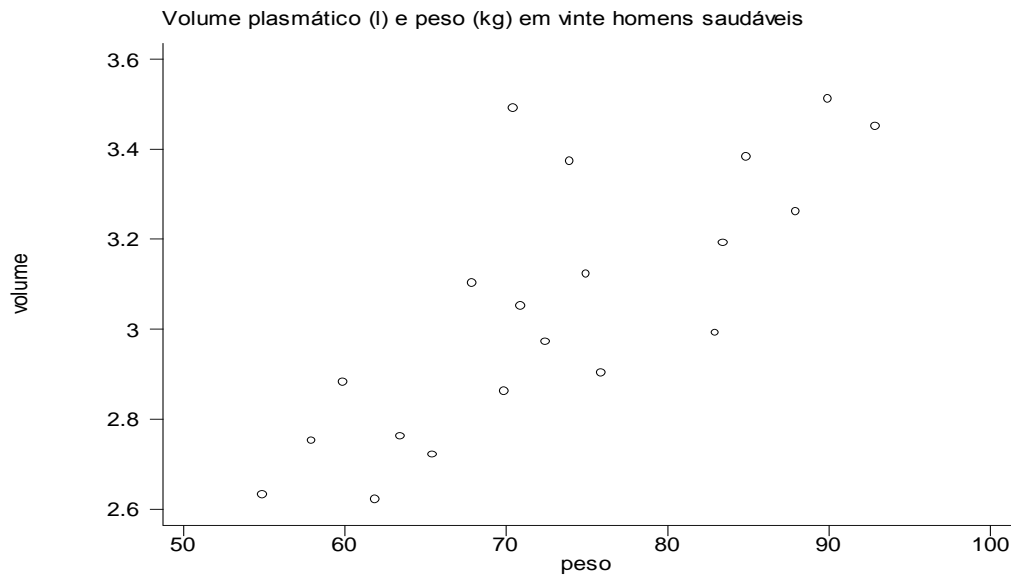
O gráfico obtido foi:



Observando o gráfico acima, você acha que existe uma correlação linear entre o volume plasmático e o peso dos vinte homens incluídos neste banco de dados?

Podemos melhorar o visual deste gráfico, por exemplo, acrescentando um título ou exibindo os valores das variáveis de acordo com uma escala.

```
graph7 volume peso, xlab(50,60,70,80,90,100) ylab(2.6, 2.8, 3.0, 3.2, 3.4, 3.6) t1(Volume plasmático (l) e peso (kg) de vinte homens saudáveis)
```



Para obter o valor do coeficiente de correlação de Pearson podemos utilizar o comando **correlate** (que pode ser abreviado como **corr**):

```
corr peso volume
```

```
. corr peso volume
(obs=20)
```

	peso	volume
peso	1.0000	
volume	0.7803	1.0000

A saída apresenta o número de sujeitos utilizados para o cálculo (obs = 20) e o coeficiente de correlação linear entre as variáveis peso e volume, isto é, $r = 0,78$.

É possível obter os coeficientes de correlação linear entre muitas variáveis contínuas do mesmo banco. Para isto, basta digitar os nomes das variáveis após o comando **corr** (por exemplo, **corr var1 var2 var3**).

Pode ser usado também o comando **pwcorr** (pairwise correlation), que produz o mesmo resultado e permite o uso da opção **sig** que apresenta o nível de significância do coeficiente de correlação apresentado.

```
pwcorr volume peso, sig
```

```
. pwcorr volume peso, sig
```

	volume	peso
volume	1.0000	
peso	0.7803	1.0000

A saída acima apresenta, abaixo do coeficiente de correlação ($r = 0,78$), o nível de significância ($p = 0,0000$).

2. Regressão linear

A regressão linear apresenta a equação da reta que melhor descreve como a variável y aumenta (ou diminui) com um aumento na variável x . A escolha de qual será a variável a ser chamada de y é importante porque, diferentemente da correlação, as duas alternativas não fornecem o mesmo resultado. A variável y é comumente denominada variável **dependente**, e x é a variável **independente** ou **explicativa**. A técnica de regressão linear permite:

- estudar a forma da relação entre x e y , e
- obter o valor esperado de y quando conhecemos apenas o valor de x .

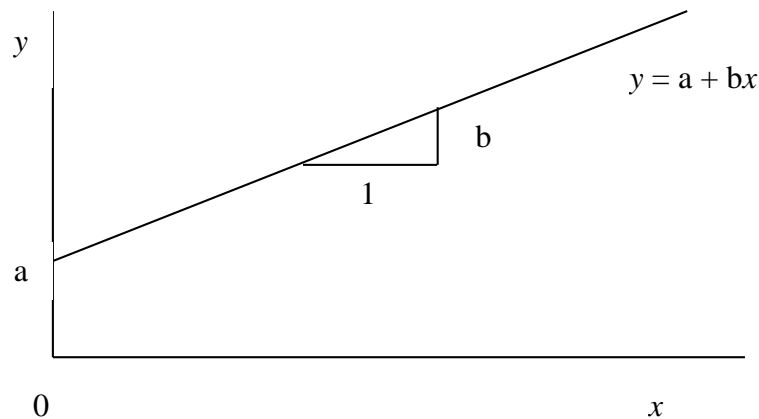
A equação da reta de regressão é:

$$y = a + bx$$

onde **a** é o intercepto e **b** é a inclinação da reta.

a (intercepto): é o ponto onde a reta cruza o eixo y e mostra o valor de y para $x=0$.

b (inclinação): mostra o aumento em y correspondente ao incremento de uma unidade em x .



Aplicação prática:

Utilizando os dados de nosso arquivo *plasma.dta* vamos utilizar a técnica de regressão linear para obter a reta que melhor exprime a relação linear entre o peso e o volume plasmático dos indivíduos incluídos no banco de dados. Nossa variável independente (x) será o peso e a variável dependente (ou resposta) será o volume plasmático (y).

Para fazer a regressão linear no STATA utilizaremos o comando `regress`. Para executarmos este comando, a variável dependente aparece em primeiro lugar, seguida da variável explicativa:

```
regress volume peso
```

Source	SS	df	MS			
Model	.967837779	1	.967837779	Number of obs =	20	
Residual	.621562203	18	.034531234	F(1, 18) =	28.03	
Total	1.58939998	19	.083652631	Prob > F =	0.0000	
				R-squared =	0.6089	
				Adj R-squared =	0.5872	
				Root MSE =	.18583	

volume	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
peso	.0204617	.003865	5.29	0.000	.0123417	.0285817
_cons	1.552716	.2858553	5.43	0.000	.9521564	2.153276

O resultado deste comando consiste em duas partes. Na primeira há uma tabela que fornece a quantidade de variação da variável **volume** explicada pelo modelo de regressão linear.

A segunda parte do resultado mostra os valores estimados para os parâmetros. O valor estimado para o parâmetro correspondente ao intercepto **a** é chamado **_cons** (constante). O valor estimado do parâmetro **b** é o coeficiente para o **peso**. Na maioria das vezes este é o parâmetro de maior interesse e pode ser chamado de coeficiente de regressão do volume plasmático com o peso.

Na saída apresentada acima, o valor estimado de a (**_cons**) é 1,55 e o valor estimado de b (peso) é 0,02.

A partir da equação geral $y = a + bx$, podemos escrever a equação de regressão utilizando as estimativas obtidas:

$$\text{volume} = 1,55 + 0,02 * (\text{peso})$$

Próximo às estimativas dos parâmetros estão os erros padrão (EP) e os correspondentes testes t e valores de p, que nos ajudam a decidir se cada parâmetro é significativamente diferente de zero. O teste para o coeficiente de regressão é o teste da hipótese nula, ou seja, de não existir relação linear. Finalmente, temos os intervalos de confiança (IC95%) dos valores dos parâmetros estimados.

Observando a saída acima, quais são os EP dos parâmetros estimados e quão forte é a evidência de que existe uma associação linear entre estas duas variáveis?

Depois de ajustar a reta de regressão, é possível calcular o volume plasmático previsto pelo modelo, dado o peso de cada indivíduo, utilizando o seguinte comando:

predict Y

O comando acima gera uma nova variável (de nome Y) onde ficam guardados os valores previstos dos volumes plasmáticos para cada peso observado. Para obter uma lista das 10 primeiras observações digite:

```
list Y peso in 1/10
```

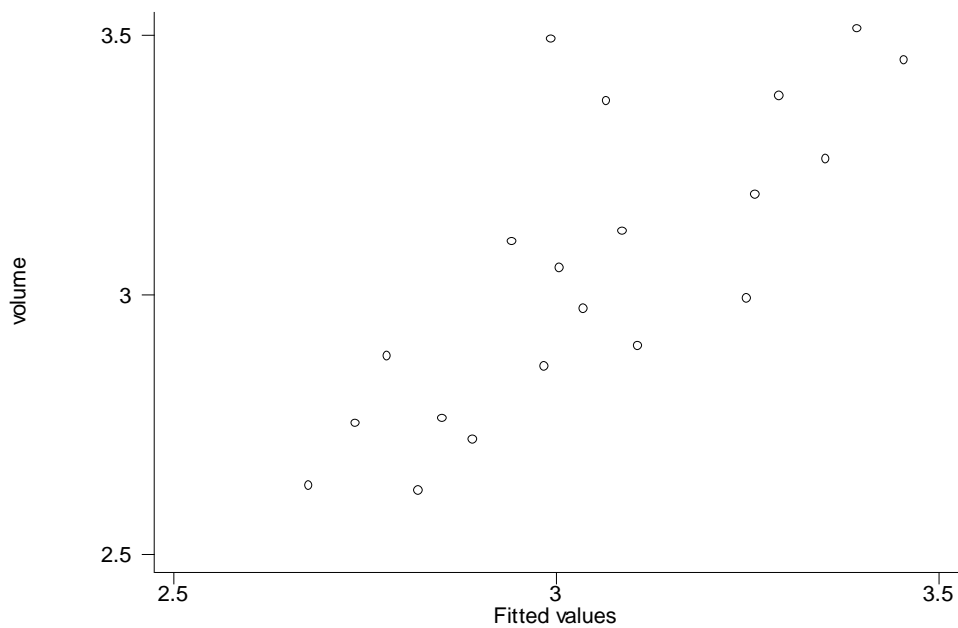
	Y	peso
1.	2.739494	58
2.	2.985034	70
3.	2.892956	65.5
4.	3.066881	74
5.	2.852033	63.5
6.	2.821341	62
7.	2.995265	70.5

8.	3.005496	71
9.	2.944111	68
10.	3.29196	85

Uma maneira descritiva de estudar a adequação do modelo adotado é desenhar um diagrama de dispersão dos valores previstos versus os valores observados:

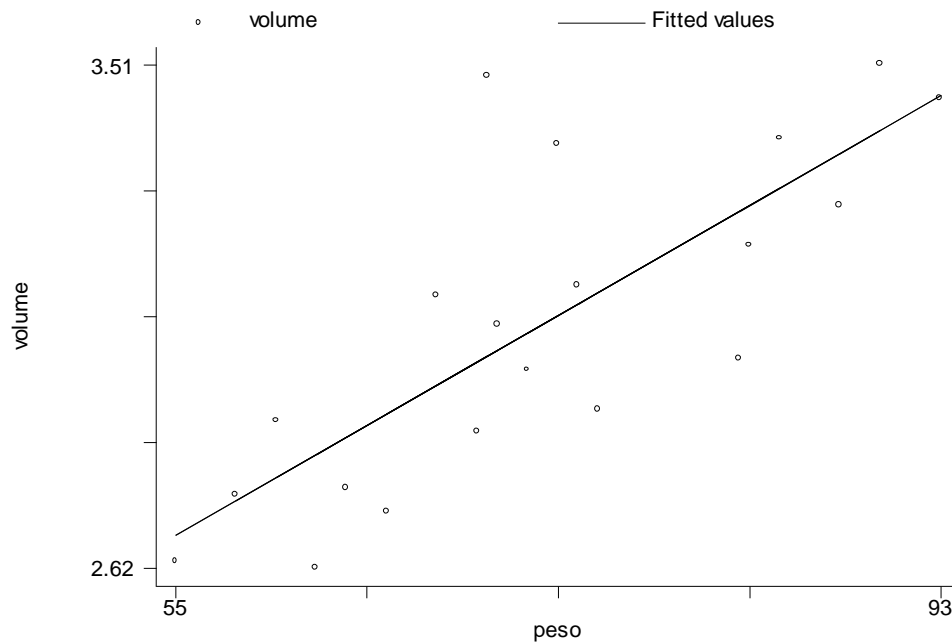
```
graph volume Y, xlabel ylabel
```

O gráfico obtido foi:



Finalmente, para construirmos o gráfico de dispersão mostrando os dados e a reta de regressão ajustada do modelo construído acima vamos utilizar o comando:

```
graph volume Y peso, c(.l) s(oi)
```



c(.l) significa “não conecte volume” e “conecte Y (valores previstos do volume)”.
s(oi) significa “use pequenos círculos para volume” e “use um símbolo invisível para Y”.

Estratificação e regressão logística

Controle de variável de confusão

O conceito de variável de confusão é central na epidemiologia moderna. De uma maneira simplificada, podemos dizer que confusão é uma "mistura de efeitos", ocorre quando a estimativa do efeito do fator de exposição estudado está misturado com o efeito de outro fator.

A variável de confusão está associada com a doença e a exposição em estudo, mas não deve estar no caminho causal da exposição para o desenvolvimento da doença.

Estratificação

Uma variável de confusão não deve ser identificada apenas através de métodos estatísticos, mas sim com base nos conhecimentos do pesquisador em relação ao problema estudado. No entanto, algumas técnicas, como estratificação e análise multivariada, podem auxiliar na identificação de uma variável de confusão.

É recomendável que a estratificação seja feita em primeiro lugar, pois é mais simples de compreender do que a análise multivariada.

Por exemplo, utilize o banco de dados **motocobr.dta** (pesquisa sobre transtornos mentais comuns – TMC- entre motoristas e cobradores de ônibus na cidade de São Paulo).

Avalie a associação entre transtorno mental comum, faixa etária e função.

Quem tem maior prevalência de TMC:

- motoristas ou cobradores?
- pessoas mais velhas ou mais novas?

Existe efeito de confundimento? O que você acha?

A variável de confusão está associada com a exposição em estudo e com o desfecho?

Para responder estas questões precisamos avaliar a associação entre faixa etária e função, entre faixa etária e TMC, e ainda, entre TMC e função nas diferentes faixas etárias.


```
tab faet fun , col row chi
```

faet	funcao		Total
	motorista	cobrador	
>29 anos	387	214	601
	64.39	35.61	100.00
	91.49	56.76	75.13
< 30 anos	36	163	199
	18.09	81.91	100.00
	8.51	43.24	24.88
Total	423	377	800
	52.88	47.13	100.00
	100.00	100.00	100.00

Pearson chi2(1) = 128.6292 Pr = 0.000

Para fazer a análise estratificada no STATA utilize o comando **mhodds**

mhodds (variável dependente) (variável de exposição) (variável de controle)

```
mhodds tmc faet
```

Maximum likelihood estimate of the odds ratio

Comparing faet==2 vs. faet==1

Odds Ratio	chi2(1)	P>chi2	[95% Conf. Interval]
1.985731	13.01	0.0003	1.357883 2.903880

```
mhodds tmc fun
```

Maximum likelihood estimate of the odds ratio

Comparing fun==1 vs. fun==0

Odds Ratio	chi2(1)	P>chi2	[95% Conf. Interval]
2.255424	19.98	0.0000	1.563236 3.254105

Análise estratificada

mhodds tmc fun faet

Mantel-Haenszel estimate of the odds ratio
Comparing fun==1 vs. fun==0, controlling for faet

Odds Ratio	chi2(1)	P>chi2	[95% Conf. Interval]	
1.910025	11.40	0.0007	1.303355	2.799080

mhodds tmc faet fun

Mantel-Haenszel estimate of the odds ratio
Comparing faet==2 vs. faet==1, controlling for fun

Odds Ratio	chi2(1)	P>chi2	[95% Conf. Interval]	
1.461327	3.55	0.0595	0.982582	2.173331

Análise Multivariada (modelo de regressão logística)

É possível examinar associações de diversas exposições com controle simultâneo de confusão

Para fazer uma regressão logística no Stata utilizamos o comando **logistic**.

logistic (variável dependente) (variável de exposição) (variável de controle)

logistic tmc fun faet

```
Logit estimates                               Number of obs   =       800
                                              LR chi2(2)      =       23.59
                                              Prob > chi2     =       0.0000
Log likelihood = -381.49553                  Pseudo R2      =       0.0300
```

tmc	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
fun	1.957635	.3925782	3.35	0.001	1.321401	2.900208
faet	1.483997	.3112098	1.88	0.060	.9838463	2.238407

Quando utilizar o comando **xi**? Quando a variável independente tem mais de duas categorias.

Em nosso exemplo:

```
. xi: logistic tmc fun faet i.sal
```

```
i.sal          _Isal_0-2          (naturally coded; _Isal_0 omitted)
Logit estimates          Number of obs   =          800
                          LR chi2(4)     =          23.97
                          Prob > chi2    =          0.0001
Log likelihood = -381.30495          Pseudo R2      =          0.0305
```

	tmc	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
fun		1.875565	.4445464	2.65	0.008	1.178638	2.984584
faet		1.45576	.3089633	1.77	0.077	.960362	2.206706
_Isal_1		1.151009	.2961827	0.55	0.585	.6950947	1.90596
_Isal_2		1.161853	.3202408	0.54	0.586	.676916	1.994193

Como você interpreta os resultados?

Análise de sobrevivência

Em situações especiais o foco de interesse do estudo é o tempo até ocorrência do evento.

A partir da amostra é possível estimar a probabilidade de um novo indivíduo sobreviver até um tempo determinado.

Para esta análise vamos utilizar o banco de dados de trabalhadores eletricitários e avaliar a probabilidade de morrer ao longo do tempo de trabalho e fatores intervenientes.

Dentro do diretório do curso utilize o banco **coorte1.dta**

Comece examinando as variáveis o banco, utilize o comando **describe**

describe

Defina as datas de entrada e saída do estudo, e o desfecho, com o comando **stset**.

```
stset (datasaid) (obito), id(id)
```

```

      id: id
      failure event: obito ~= 0 & obito ~= .
obs. time interval: (datasaid[_n-1], datasaid]
      exit on or before: failure

```

```

-----
-
10017 total obs.
   0 exclusions
-----
-
10017 obs. remaining, representing
10017 subjects
  498 failures in single failure-per-subject data
1.26e+08 total analysis time at risk, at risk from t = 0
                                     earliest observed entry t = 0
                                     last observed exit t = 12783

```

Veja como o STATA produz os tempos de observação

stdes

```

      failure _d: obito
      analysis time _t: datasaid
      id: id

```

Category	total	per subject			
		mean	min	median	max
no. of subjects	10017				
no. of records	10017	1	1	1	1
(first) entry time		0	0	0	0
(final) exit time		12587.51	6940	12783	12783
subjects with gap	0				
time on gap if gap	0
time at risk	1.261e+08	12587.51	6940	12783	12783
failures	498	.0497155	0	0	1

Calcule a taxa de mortalidade e interprete o resultado.

strate

```

      failure _d: obito
      analysis time _t: datasaid
      id: id

```

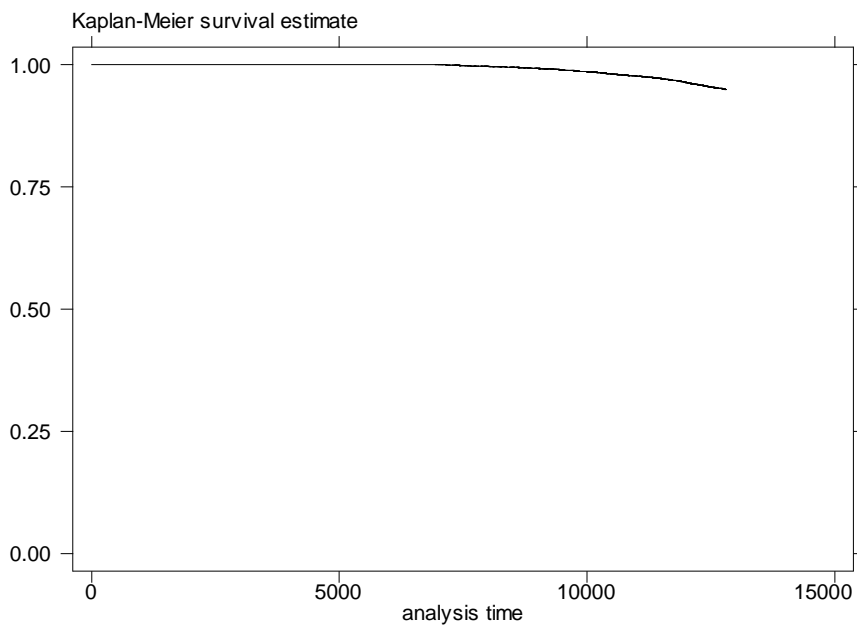
Estimated rates and lower/upper bounds of 95% confidence intervals
(10017 records included in the analysis)

<u>D</u>	<u>Y</u>	<u>Rate</u>	<u>Lower</u>	<u>Upper</u>
498	126089094.000	0.0000039	0.0000036	0.0000043

É possível visualizar graficamente a curva de sobrevivência da coorte.
Pode-se também obter a curva inversa, de incidência (mortalidade).

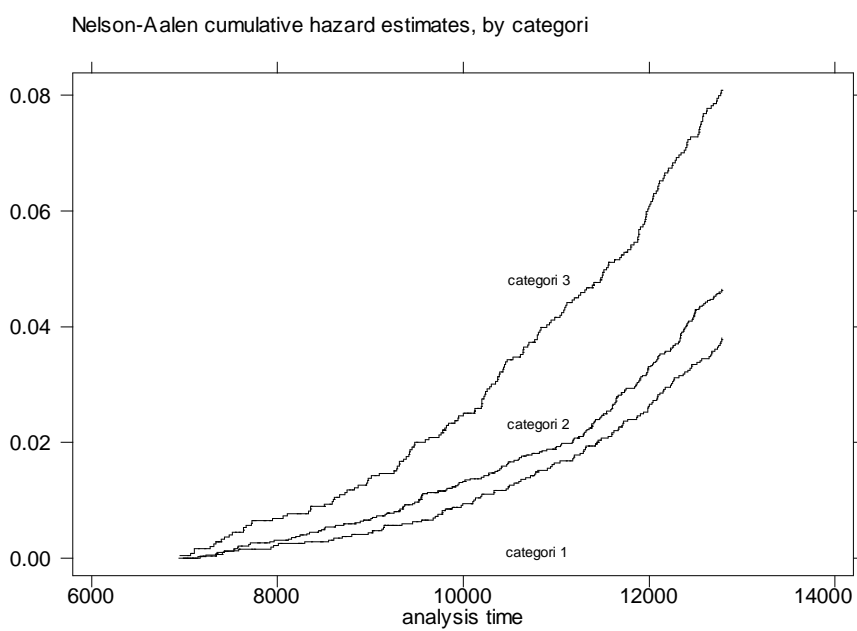
Tente interpretar as curvas obtidas.

sts



Agora verifique se a mortalidade é diferente segundo categoria ocupacional.

sts graph, na by (category)



sts test categori, logrank

```
failure _d: obito
analysis time _t: datasaid
id: id
```

Log-rank test for equality of survivor functions

categori	Events observed	Events expected
1	118	161.94
2	191	214.51
3	189	121.55
Total	498	498.00

chi2(2) = 51.93
Pr>chi2 = 0.0000

Outros recursos do STATA

O STATA apresenta vários recursos de ajuda ao usuário. Além do manual impresso (composto por 7 volumes), outros recursos disponíveis são:

Comando HELP (exibido na tela de apresentação do STATA):

Clicando no item “*help*” aparecerá uma caixa com vários comandos: o comando “*search*” permite que ao digitarmos um tópico sobre o qual necessitamos de informação sejam localizados todos os comandos do STATA relacionados ao tema.

Suponha que você deseja saber como calcular intervalos de confiança (*confidence intervals*) utilizando o STATA. Clique em HELP, em seguida SEARCH e, na caixa de diálogo que surgir digite “*confidence interval*”. O STATA exibirá a seguinte tela:

```
Chapter 23 . . . . . Estimation and post-estimation commands
(help est, postest)

[R] bitest . . . . . Binomial probability test
(help bitest)

[R] centile . . . . . Report centile and confidence interval
(help centile)
```

- [R] `ci` Confidence intervals for means, proportions, and counts
(help [ci](#))

- [R] `epitab` Tables for epidemiologists
(help [epitab](#))

- [R] `level` Set default confidence level
(help [level](#))

- [R] `lincom` Linear combinations of estimators
(help [lincom](#))

- [R] `ltable` Life tables for survival data
(help [ltable](#))

Estes são alguns comandos em que são utilizados ou calculados intervalos de confiança. Observe que abaixo do comando aparece entre parênteses help “[nome do comando](#)”. Este é o link para a ajuda de cada comando listado. Experimente clicar sobre um deles.

Em outras situações, conhecemos o nome do programa, mas temos dúvidas sobre como utilizá-lo ou sobre as opções que ele permite. Suponha que você sabe que o comando para calcular intervalos de confiança é o comando `ci`, mas você tem dúvidas sobre ele. Clique em HELP, em seguida STATA COMMAND. Na caixa de diálogo que aparecerá digite: “`ci`”. O STATA exibirá a tela:

```

help for ci, cii                                     (manual: [R] ci)
-----
Confidence intervals for means, proportions, and counts

          ci [varlist] [weight] [if exp] [in range] [, level(#) binomial
            poisson exposure(varname) total ]

(normal)   cii #obs      #mean #sd [, level(#) ]
(binomial) cii #obs      #succ   [, level(#) ]
(Poisson)  cii #exposure #events  , poisson [ level(#) ]

by ... : may be used with ci (but not with cii); see help by.

aweight and fweight are allowed; see help weights.

Description

ci computes standard errors and confidence intervals for each of the variables in
varlist. cii is the immediate form of ci; see help immed.

Options

level(#) specifies the confidence level, in percent, for confidence intervals;
see help level.

binomial tells ci that the variables are 0/1 binomial variables; exact binomial

```

confidence intervals are calculated. (cii produces binomial confidence intervals when only two numbers are specified.)

poisson specifies that the variables (or numbers in the case of cii) are Poisson-distributed counts; exact confidence intervals will be calculated.

exposure(varname) is used only with poisson. It is not necessary to specify poisson if exposure() is specified. varname contains the total exposure (typically a time or an area) during which the number of events recorded in varlist were observed.

total is for use with the by ... : prefix; it requests that, in addition to ouput for each by-group, output be added for all groups combined.

Examples

```
. ci mpg
. ci mpg price, level(90)
. ci promoted, binomial
. ci count, poisson
. ci deaths, exposure(pyears)
. cii 166 19509 4379                      (166 obs, mean=19509, sd=4379)
. cii 166 19509 4379, level(90)
. cii 10 1                                (10 binomial events, 1 observed success)
. cii 1 27, poisson                       (27 Poisson events observed)
```

Also see

Manual: [R] ci

On-line: help for bitest, centile, immed, means, pctile, summarize, ttest

A tela de ajuda de comando apresenta a descrição, as opções, exemplos de utilização e outros comandos relacionados.

O HELP também oferece links para o site oficial do STATA. Uma das ferramentas mais úteis para os usuários que estão se familiarizando com o STATA são as “FREQUENTLY ASKED QUESTIONS” (dúvidas mais freqüentes). Para utilizar este recurso, clique em HELP, em seguida STATA WEB SITE e em seguida FREQUENTLY ASKED QUESTIONS (FAQ) (este recurso só pode ser utilizado se você estiver conectado à internet). Há um recurso de busca na página de FAQs e também são exibidos os temas já abordados, agrupados por assunto (estatística, gráficos etc).

O site do STATA (que pode ser acessado a partir do HELP ou diretamente através do endereço: <http://www.stata.com>) oferece outros recursos de suporte ao usuário, tais como informações sobre cursos on-line, livros, publicações e atualizações do programa.

Leitura recomendada

Altman, D. E. (1991). **Practical Statistics for Medical Research**. London, Chapman & Hall.

Berquó, E. S.; Souza, J. M. P.; Gotlieb, S. L. D. (1981). **Bioestatística**. São Paulo, Editora Pedagógica Universitária.

Hosmer D.W. e Lemeshow S. (1989). **Applied logistic regression**. John Wiley & Sons.

Kirkwood, B. R. (1988). **Essentials of Medical Statistics**. Oxford, Blackwell Science Publications.

Kleinbaum, D. G. (1996). **Survival Analysis**. Springer-Verlag, New York.

Souza, M.F.M. (1996). **Um estudo sobre o risco de distúrbios psiquiátricos menores entre motoristas e cobradores do sistema de ônibus urbano na cidade de São Paulo**. Tese de mestrado. Faculdade de Medicina - USP.