

# STATA introduction course

## Solutions to Exercises

```
*****  
*                                                                 *  
* EXERCISE 1                                                                 *  
*                                                                 *  
*****
```

```
. use numberofwords,clear
```

### ANSWER TO QUES 1

```
. tab age
```

age	Freq.	Percent	Cum.
16	91	33.96	33.96
18	79	29.48	63.43
20	98	36.57	100.00
Total	268	100.00	

### ANSWER TO QUES 2

```
. tab age if region=="rural"
```

age	Freq.	Percent	Cum.
16	53	33.76	33.76
18	49	31.21	64.97
20	55	35.03	100.00
Total	157	100.00	

### ANSWER TO QUES 3

```
. tab age region, col
```

```
+-----+  
| Key |  
+-----+  
| frequency |  
| column percentage |  
+-----+  
  
      |      region  
      |      rural      urban |      Total  
-----+-----+-----+  
      |      53      38 |      91
```

	33.76	34.23	33.96
18	49	30	79
	31.21	27.03	29.48
20	55	43	98
	35.03	38.74	36.57
Total	157	111	268
	100.00	100.00	100.00

**or**

```
. tab region age, row
```

```
+-----+
| Key      |
+-----+
| frequency|
| row percentage|
+-----+
```

region	age			Total
	16	18	20	
rural	53	49	55	157
	33.76	31.21	35.03	100.00
urban	38	30	43	111
	34.23	27.03	38.74	100.00
Total	91	79	98	268
	33.96	29.48	36.57	100.00

The percentages are quite similar between the children belonging to urban and rural regions.

### **ANSWER TO QUES 4**

We can look at the dataset raw in the following manner:

```
. use raw, clear
```

```
. list in 1/10
```

```
+-----+
| id  age  region  gender  nwords  wordst~t |
+-----+
1. | 1    9   urban    1      0      0 |
2. | 1   10   urban    1      0      0 |
3. | 1   11   urban    1      0      0 |
4. | 1   12   urban    1      0      0 |
5. | 1   13   urban    1      0      0 |
+-----+
6. | 1   14   urban    1      0      0 |
```

7.		1	15	urban	1	0	0	
8.		1	16	urban	1	0	0	
9.		1	17	urban	1	0	0	
10.		1	18	urban	1	6	1	

+-----+

This is the raw data set, where we have several records for each child and each record is for each month of the child and `nwords` represents the number of words spoken by that age and `wordstart` denotes whether the child had started to speak or not.

. codebook

-----  
group(id)  
-----

```

type: numeric (float)
range: [1,10]                units: 1
unique values: 10            missing ..: 0/120

mean: 5.5
std. dev: 2.88432

percentiles:    10%    25%    50%    75%    90%
                1.5    3     5.5    8     9.5

```

-----  
age  
(unlabeled)  
-----

```

type: numeric (byte)
range: [9,20]                units: 1
unique values: 12            missing ..: 0/120

mean: 14.5
std. dev: 3.46653

percentiles:    10%    25%    50%    75%    90%
                10    11.5  14.5  17.5  19

```

-----  
region  
(unlabeled)  
-----

```

type: string (str5)
unique values: 2              missing "": 0/120

tabulation:  Freq.  Value
              36    "rural"
              84    "urban"

```

-----  
gender  
(unlabeled)  
-----

```

type: numeric (float)
range: [1,2] units: 1
unique values: 2 missing .: 0/120

tabulation: Freq. Value
              72  1
              48  2

```

-----  
nwords  
(unlabeled)  
-----

```

type: numeric (float)
range: [0,540] units: 1
unique values: 33 missing .: 0/120

mean: 27.9
std. dev: 83.0409

percentiles: 10% 25% 50% 75% 90%
              0  0  0  19  57

```

-----  
wordstart  
(unlabeled)  
-----

```

type: numeric (float)
range: [0,1] units: 1
unique values: 2 missing .: 0/120

tabulation: Freq. Value
              77  0
              43  1

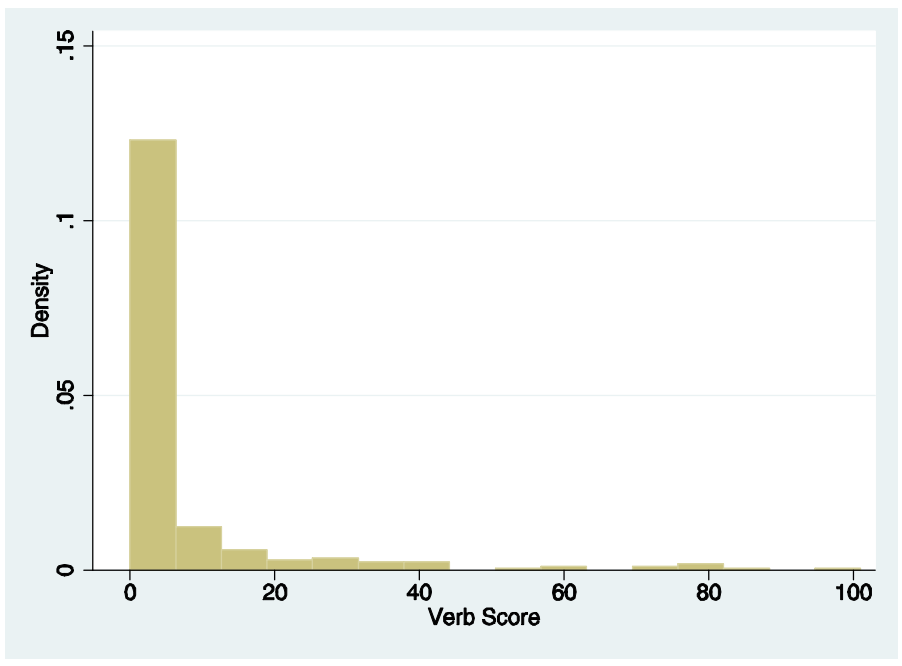
```

```
*****
*
* EXERCISE 3
*
*****
```

```
. use numberofwords, clear
```

**ANSWER TO QUES 1**

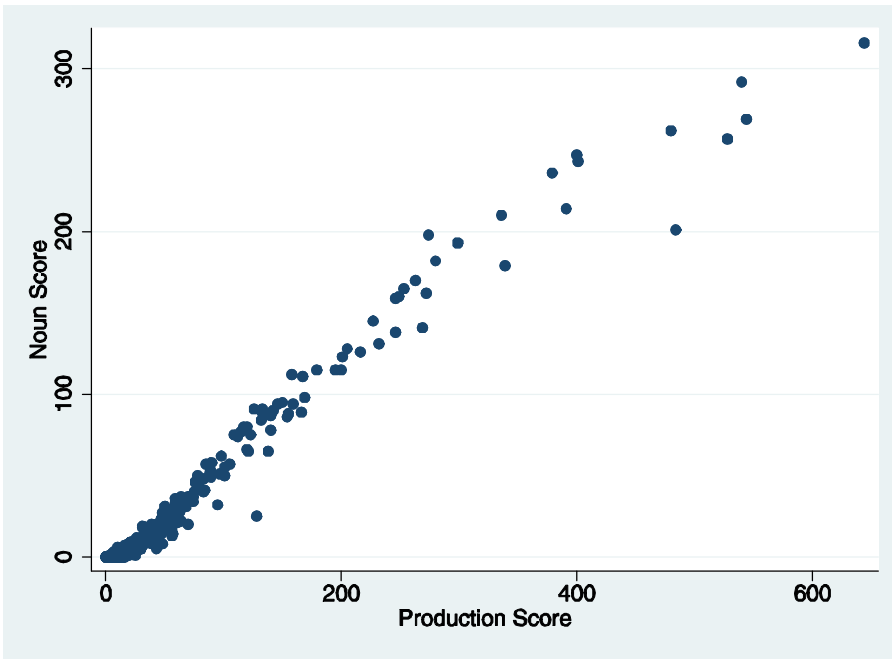
```
. histogram verbs
(bin=16, start=0, width=6.3125)
```



The graph shows that the distribution of the number of verbs spoken are skewed to the left and there were very few children who knew more than 20 verbs.

**ANSWER TO QUES 2**

```
. scatter nouns words
```



We can see that the number of nouns is highly correlated with the number of words spoken.

**ANSWER TO QUES 3**

```
. scatter verbs nouns, by(gender)
```



From the graph we can see that the relationship of nouns and verbs are not very different between boys and girls.

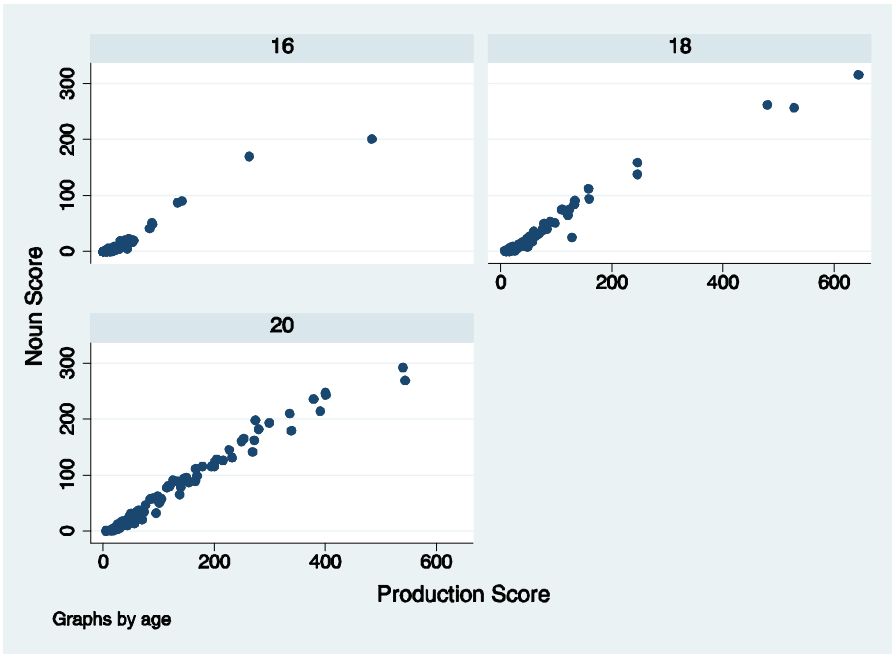
```

*****
*
* EXERCISE 4
*
*****

```

```
. use numberofwords, clear
```

```
. sc n w, by(age)
```



```

*****
*
* EXERCISE 5
*
*****

```

*ANSWER TO QUES 1*

```

. use numberofwords, clear
. gen fast=words>400
. tab gender fast, row

```

```

+-----+
| Key          |
+-----+
| frequency    |
| row percentage |
+-----+

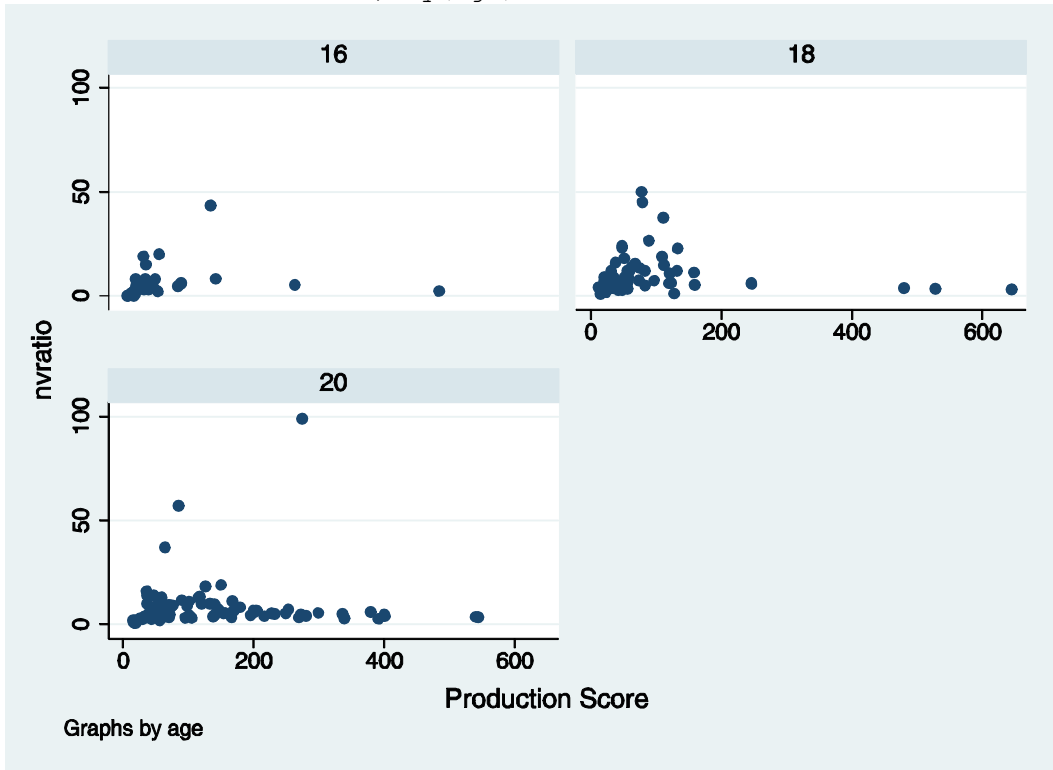
```

gender	fast		Total
	0	1	
male	120	5	125
	96.00	4.00	100.00
female	141	2	143
	98.60	1.40	100.00
Total	261	7	268
	97.39	2.61	100.00

We can find a difference between boys and girls with respect to being fast because 4% of the boys have a vocabulary more than 400 words while among girls, it is only 1.4%.

ANSWER TO QUES 2

```
. gen nvratio=noun/verb  
(96 missing values generated)  
  
. scatter nvratio words, by(age)
```



Though the noun/verb ratio seems to increase slightly with the vocabulary till a vocabulary of about 200 words, it remains low and constant beyond that.

```
*****
*
* EXERCISE 6
*
*****
```

**ANSWER TO QUES 1**

```
. help keep
help drop, help keep          tool:  Variables Manager
observations                  dialog: keep/drop
-----
```

Title

[D] drop -- Eliminate variables or observations

Syntax

Drop variables

drop varlist

Drop observations

drop if exp

Drop a range of observations

drop in range [if exp]

Keep variables

keep varlist

Keep observations that satisfy specified condition

keep if exp

Keep a range of observations

keep in range [if exp]

by is allowed with the second syntax of drop and the second syntax of keep; see [D] by.

Menu

Keep or drop variables

Data > Variables Manager

## Keep or drop observations

Data > Create or change data > Keep or drop observations

### Description

drop eliminates variables or observations from the data in memory.

keep works the same as drop, except that you specify the variables or observations to be kept rather than the variables or observations to be deleted.

Warning: drop and keep are not reversible. Once you have eliminated observations, you cannot read them back in again. You would need to go back to the original dataset and read it in again. Instead of applying drop or keep for a subset analysis, consider using if or in to select subsets temporarily. This is usually the best strategy. Alternatively, applying preserve followed in due course by restore may be a good approach.

### Examples

#### Setup

```
. sysuse census
```

#### Describe the data

```
. describe
```

#### Drop all variables with names that begin with pop

```
. drop pop*
```

#### Describe the resulting data

```
. describe
```

#### Drop marriage and divorce

```
. drop marriage divorce
```

#### Describe the resulting data

```
. describe
```

#### Drop any observation for which medage is greater than 32

```
. drop if medage > 32
```

#### Drop the first observation for each region

```
. by region, sort: drop if _n == 1
```

#### Drop all but the last observation in each region

```
. by region: drop if _n != _N
```

#### Keep the first 2 observations in the dataset

```
. keep in 1/2
```

#### Describe the resulting data

```
. describe
```

#### Drop all observations and variables

```
. drop _all
```

#### Describe the resulting data

```
. describe
```

Also see

Manual: [D] drop

Help: [D] clear, [D] varmanage

## **ANSWER TO QUES 2**

. search pearson

Keyword search

Keywords: pearson

Search: (1) Official help files, FAQs, Examples, SJs, and STBs

Search of official help files, FAQs, Examples, SJs, and STBs

- [R] correlate . . Correlations (covariances) of variables or coefficients  
(help correlate)
- [R] logistic . . . . . Logistic regression, reporting odds ratios  
(help logistic)
- [R] logistic postestimation . . . . . Postestimation tools for logistic  
(help logistic postestimation)
- [MV] cluster programming utilities . Cluster-analysis programming utilities  
(help cluster programming)
- [MV] matrix dissimilarity . . Compute similarity or dissimilarity measures  
(help matrix dissimilarity)
- [MV] measure\_option . . . Option for similarity and dissimilarity measures  
(help measure\_option)
- [P] matrix dissimilarity . . Compute similarity or dissimilarity measures  
(help matrix dissimilarity)
- FAQ . . . . . A comparison of different tests for trend  
. . . . . W. Sribney  
3/96 Does Stata provide a test for trend?  
<http://www.stata.com/support/faqs/stat/trend.html>
- Example . . Regression Models for Categorical and Limited Dependent Variables  
. . . . . UCLA Academic Technology Services  
2/08 examples from the book Regression Models for Categorical  
and Limited Dependent Variable by J. Scott Long  
<http://www.ats.ucla.edu/stat/stata/examples/long/>
- Example . . . . . Textbook examples: An Introduction to Categorical Analysis  
. . . . . UCLA Academic Technology Services  
9/07 examples from the book An Introduction to  
Categorical Analysis by Alan Agresti  
<http://www.ats.ucla.edu/stat/stata/examples/icda/>
- Example . . Textbook examples: Regression Analysis by Example, Third Edition  
. . . . . UCLA Academic Technology Services  
7/06 examples from the book Regression Analysis by  
Example, Third Edition by Samprit Chatterjee,  
Ali S. Hadi & Bertram Price

<http://www.ats.ucla.edu/stat/stata/examples/chp/>

- Example . . Textbook examples: Computer-Aided Multivariate Analysis, Fourth Ed  
. . . . . UCLA Academic Technology Services  
10/05 examples from the book Computer-Aided Multivariate  
Analysis, 4th Edition by A. A. Afifi, V. Clark and  
S. May  
<http://www.ats.ucla.edu/stat/stata/examples/cama4/>
- SJ-8-3 pr0041 . Speaking Stata: Corr. with confidence, Fisher's z revisited  
(help corrci, corrcii if installed) . . . . . N. J. Cox  
Q3/08 SJ 8(3):413--439  
reviews Fisher's z transformation and its inverse, the  
hyperbolic tangent, and reviews their use in inference  
with correlations
- SJ-8-2 st0142 . Multinomial GOF: Large-sample svy & small-sample exact tests  
(help mgof if installed) . . . . . B. Jann  
Q2/08 SJ 8(2):147--169  
computes distributional tests for discrete (categorical,  
multinomial) variables and supports large-sample tests for  
complex survey designs and exact tests for small samples
- SJ-6-1 st0099 . . GOF test for logistic reg. fitted using survey sample data  
(help svylogitgof if installed) . . . . . K. J. Archer and S. Lemeshow  
Q1/06 SJ 6(1):97--105  
estimates the F-adjusted mean residual test after svy: logit  
or svy: logistic
- SJ-5-4 st0092 Extended GLM: simultaneous est. of flex. link & variance func.  
(help pglm if installed) . . . . . A. Basu  
Q4/05 SJ 5(4):501--516  
simultaneously solves the extended estimating equations  
estimator for parameters in the link and variance functions  
along with those of the linear predictor in a generalized  
linear model (GLM)
- SJ-4-4 sg159\_1 . . . . . Software update for ci2 and cii2  
(help ci2 if installed) . . . . . P. T. Seed  
Q4/04 SJ 4(4):490  
ci2 and cii2 updated to work with Stata 7.0 and later;  
bug fix for cii2
- SJ-3-1 st0031 . . Intra-class corr. in random-effects models for binary data  
(help xtrho, xtrhoi if installed) . . . . . G. Rodriguez & I. Elo  
Q1/03 SJ 3(1):32--46  
intra-class manifest associations in random-effects xtprobit,  
xtlogit, and xtclog including Pearson's r and Yule's Q
- STB-59 sg159 . . . . . Confidence intervals for correlations  
(help ci2 if installed) . . . . . P. T. Seed  
1/01 pp.27--28; STB Reprints Vol 10, pp.267--269  
enhancement of ci and cii that provide confidence intervals  
for Pearson's product moment correlation and Spearman's  
rank correlation
- STB-55 sg138 . . . . . Bootstrap inferences about measures of correlation  
(help bootcor if installed) . . . . . D. J. Neal  
5/00 pp.17--20; STB Reprints Vol 10, pp.149--152  
provides bootstrap confidence interval and test for the  
difference of two Pearson's R, intraclass correlations, or  
concordance coefficients

STB-51 sg118 . . Partitions of chi-squared for ordered column two-way tables  
(help opartchi if installed) . . . . . R. Wolfe  
9/99 pp.37--40; STB Reprints Vol 9, pp.203--207  
partitions Pearson's chi-squared statistic into components  
that describe row differences for two-way tables with ordered  
columns and provides a test of trend

STB-49 sg64.1 . . . . . Update to pwcorr  
(help pwcorr if installed) . . . . . F. Wolfe  
5/99 p.17; STB Reprints Vol 9, p.159  
Fix to pwcorr

STB-35 sg64 . . . . . pwcorr: An enhanced correlation display  
(help pwcorr if installed) . . . . . F. Wolfe  
1/97 pp.22--25; STB Reprints Vol 6, pp.163--167  
combines pwcorr and spearman into single command that has  
enhanced formatting

(end of search)

**A search on pearson gives many results. You can narrow your search by specifying two keywords**

. search pearson confidence

Keyword search

Keywords: pearson confidence  
Search: (1) Official help files, FAQs, Examples, SJs, and STBs

Search of official help files, FAQs, Examples, SJs, and STBs

SJ-8-3 pr0041 . Speaking Stata: Corr. with confidence, Fisher's z revisited  
(help corrci, corrcii if installed) . . . . . N. J. Cox  
Q3/08 SJ 8(3):413--439  
reviews Fisher's z transformation and its inverse, the  
hyperbolic tangent, and reviews their use in inference  
with correlations

SJ-5-4 st0092 Extended GLM: simultaneous est. of flex. link & variance func.  
(help pglm if installed) . . . . . A. Basu  
Q4/05 SJ 5(4):501--516  
simultaneously solves the extended estimating equations  
estimator for parameters in the link and variance functions  
along with those of the linear predictor in a generalized  
linear model (GLM)

SJ-4-4 sg159\_1 . . . . . Software update for ci2 and cii2  
(help ci2 if installed) . . . . . P. T. Seed  
Q4/04 SJ 4(4):490  
ci2 and cii2 updated to work with Stata 7.0 and later;  
bug fix for cii2

STB-59 sg159 . . . . . Confidence intervals for correlations  
(help ci2 if installed) . . . . . P. T. Seed  
1/01 pp.27--28; STB Reprints Vol 10, pp.267--269  
enhancement of ci and cii that provide confidence intervals  
for Pearson's product moment correlation and Spearman's  
rank correlation

```
STB-55  sg138 . . . . . Bootstrap inferences about measures of correlation
        (help bootcor if installed) . . . . . D. J. Neal
        5/00  pp.17--20; STB Reprints Vol 10, pp.149--152
        provides bootstrap confidence interval and test for the
        difference of two Pearson's R, intraclass correlations, or
        concordance coefficients
```

(end of search)

You can see that STB-59 and SJ-4-4 have exactly what we are looking for. Now you can use `findit` to download it.

```
. findit pearson confidence
```

If you downloaded `ci2` by clicking on `sg159` in STB-59 and got it installed, you can type

```
. help ci2
. use numberofwords, clear
. version 7
. ci2 words nouns, corr
```

```
Confidence interval for Pearson's product-moment correlation
of words and nouns, based on Fisher's transformation.
Correlation = 0.984 on 268 observations (95% CI: 0.979 to 0.987)
```

```
. version 11
```

The "version 7" is necessary here, because `ci2` was written under version 7 and seems to be not compatible with version 11. After the exercise is over, type "version 11" to revert back to version 11.

The better choice would be to use SJ-4-4 where the command has been updated to be compatible with the later versions. Click on `sg159_1` and you can directly use the command

```
. help ci2
. use numberofwords, clear
. ci2 words nouns, corr
```

```
Confidence interval for Pearson's product-moment correlation
of words and nouns, based on Fisher's transformation.
Correlation = 0.984 on 268 observations (95% CI: 0.979 to 0.987)
```

```
*****  
*                                                                 *  
* EXERCISE 7                                                                 *  
*                                                                 *  
*****
```

*ANSWER TO QUES 1*

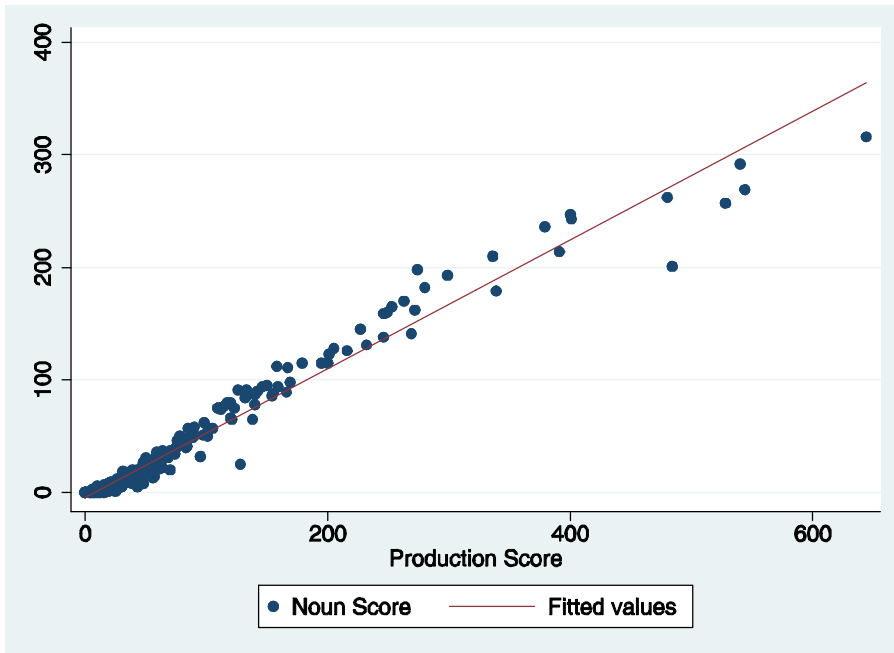
You can open the do-file editor and try to type in the following lines:

```
log using exercise7, text replace  
use numberofwords, clear  
tab age gender, row  
log close
```

```
*****  
*                                                                 *  
* EXERCISE 8                                                                 *  
*                                                                 *  
*****
```

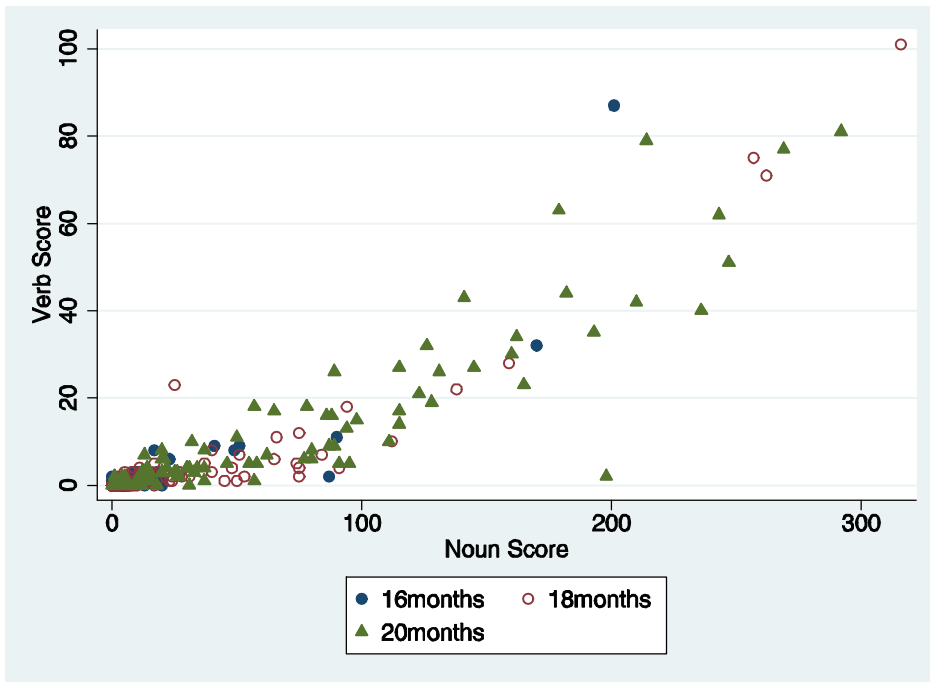
ANSWER TO QUES 1

```
. use numberofwords, clear  
  
. scatter nouns words ||lfit nouns words
```



**ANSWER TO QUES 2**

```
. scatter verbs nouns if age==16, msym(O) || scatter verbs nouns if age==18,  
msym(Oh) || scatter verbs nouns if age==20, msym(T) leg(lab(1 "16months") lab(2 "18months")  
lab(3 "20months"))
```



```

*****
*
* EXERCISE 9
*
*****

```

ANSWER TO QUES 1

```

. use numberofwords, clear

. gen starttalk=words>0

. gen agefw=ageat1w
(10 missing values generated)

. replace agefw=age if agefw==.
(10 real changes made)

. stset agefw, fail(starttalk==1)

      failure event:  starttalk == 1
obs. time interval:  (0, agefw]
exit on or before:  failure

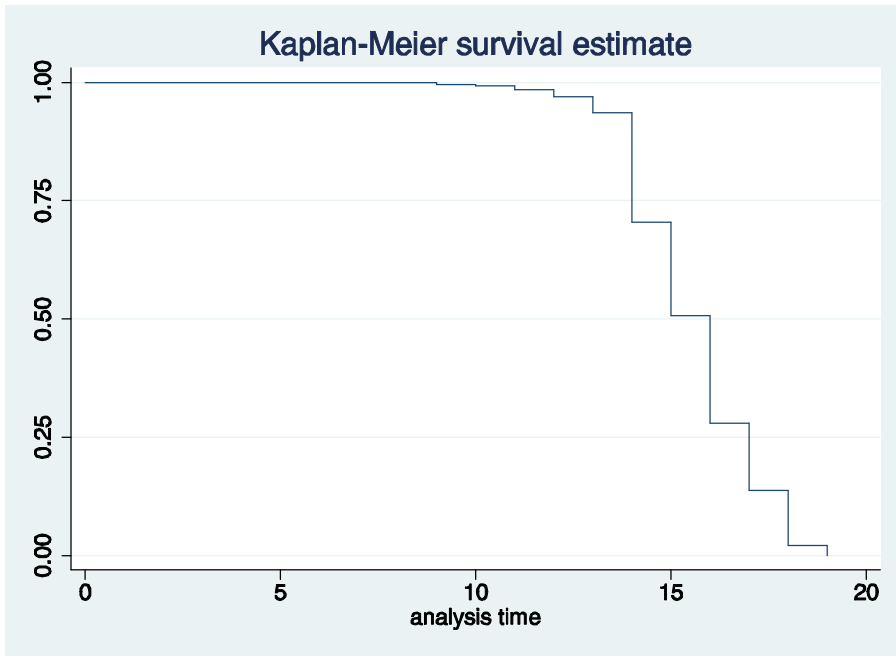
-----
      268  total obs.
       0  exclusions
-----

      268  obs. remaining, representing
      258  failures in single record/single failure data
      4147 total analysis time at risk, at risk from t =           0
              earliest observed entry t =           0
              last observed exit t =           19

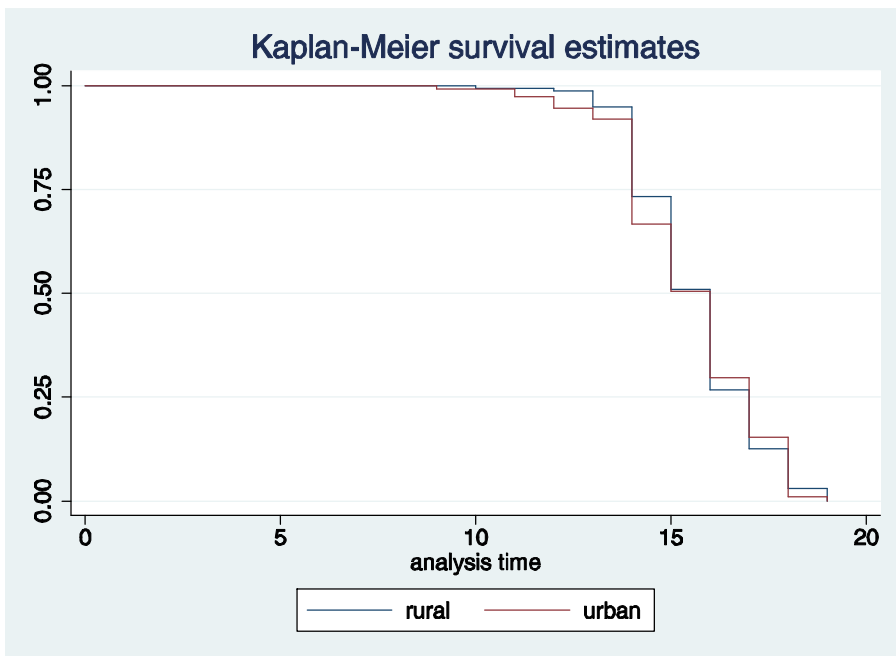
. sts graph

      failure _d:  starttalk
analysis time _t:  agefw

```



```
. sts graph, by(region)
      failure _d: starttalk
      analysis time _t: agefw
```



```
. sts test region
      failure _d: starttalk
      analysis time _t: agefw
```

Log-rank test for equality of survivor functions

region	Events observed	Events expected
rural		
urban		

rural		151	151.63
urban		107	106.37
Total		258	258.00
		chi2 (1) =	0.01
		Pr>chi2 =	0.9181

From the Kaplan Meier graph, we could see that the time to first word did not show a different pattern between the children from the rural and the urban regions. And the log rank test shows that the difference in the distribution is not statistically significant at 5% level of significance.