

Shazam Notes (R. Butler, adapted from the Shazam Manual)
WRITING YOUR PAPER WITH SHAZAM

A SIMPLE PROGRAM AND FEW BASIC PROCEDURES IN SHAZAM

Non-executable **comment statements** (optional, to remind us why we did what we did or where the data is located or when we did the analysis, etc.) starts with an ‘*’ in the first column, as in the sample program below. If you want to continue the comment to the next line, or any of the commands to the next line, just put a ‘&’ as your last right hand side character and the program will continue reading—as part of the current line—whatever is on the next line. An example of a two-lined comment statement would be:

```
*This is a program for Coach Butler, and I love to take his courses &  
because he has a 4-year warranty, which is more than many Presidents
```

The **sample** statement is not optional, and indicates which observations are to be used in the analysis. In the example below, all 20 of the sample points are to be read in and used, as indicated by ‘sample 1 20’ (which means, start reading at observation 1 and continue until you reach observation 20). The **read** statement is not optional either, it tells the program the name of the variables (in the columns) that are being read in; the read statement also indicates which file to use (if no file in parenthesis follows read it assumes the data will be found on the first line after the read statement), so that ‘read(utah_cps.txt) wage age race sex’ means read in the data from the utah_cps.txt file and the variable in the first column is wage, the second column is age, the third column is ‘race’ and the fourth column contains the values for the ‘sex’ variable. After the data is read in, it can be manipulated (no manipulations are contained in the simple program below) and then various procedures invoked to do the statistical analysis. A simple program with three such procedures is:

```
*sample program to illustrate shazam's basic procedures  
sample 1 20  
read wage college age male married  
375 1 28 1 1  
176 1 21 1 0  
64 0 63 1 0  
934 1 54 1 1  
877 0 57 1 1  
874 1 62 1 0  
748 1 46 1 0  
375 0 28 1 0  
129 0 23 1 1  
790 1 57 0 0  
971 1 73 0 1  
773 0 52 0 1  
664 1 29 0 0  
675 1 44 0 1  
556 0 58 0 1  
387 0 26 0 1  
278 0 29 0 1  
419 1 33 0 1  
670 1 42 0 1
```

```

371 0 31 0 0
* "stat" does descriptive stats for indicated var
stat wage college age male married
* "ols" does regression with first var=dep. var
ols wage college age male
* "logit" does qualitative dep var regression
logit married college age male
stop

```

Some of the more useful procedures you may need to know about include the following:

STAT [var1 var2 var3 ...] – calculates the basic descriptive statistics for the variables

OLS [dep. var.] [indep. var.] [/option]– regresses the dependent variables on the specified number of independent variables. Some useful options, that come after last of the independent variables, includes:

/hetcov – corrects for heteroskedasticity

/pred=[name] –saves the predicted value of the dependent variable from the regression and gives it the '[name]' you specify for subsequent use in the program

/resid=[name] –does the same for the regression residuals

/weight=corrects for heteroskedasticity using the weight you specify (multiplies all the terms by '1/square root of the weight.'

DIAGNOS/HET—checks for heteroskedasticity, comes immediately after the OLS command

LOGIT [dep. var.] [indep. var.] –runs the logit regression model where the dependent variables is binary (a yes or no-type) variable

TEST {[var]=[value]}— this is useful for testing multiple regressions on the regression coefficients at the same time. In order to test if the age and male coefficients are jointly zero in the above example, we would modify the program by adding the following after the OLS command (so including the OLS command, it would look like):

```

ols wage college age male
test
test age=0
test male=0
end

```

STOP or QUIT—indicates the end of your program

MORE COMPLEXITY: MANIPULATING DATA and the FILE PATH command

The data that you read down from the web will be too large to include inside your program, and it will contain some variables that will have to be manipulated. These commands are contained in the following program:

```

*utahcps3.sha reads utah_cps.asc downloaded from the ferret website
file path c:\my2000docs\BYU classes\econ388\classrm_data\
sample 1 1110
read(utah_cps.asc) wklywg age race sex stateid
if (wklywg.eq.0) wklywg=-99999
set skipmiss

```

```

if (sex.eq.1) male=1
if (sex.eq.2) male=0
if (race.eq.1) white=1
if (race.ne.1) white=0
skipif (wklywg.le.0)
genr lnwage = log(wklywg)
ols lnwage age white male
end
stop

```

The first new type of statement encountered in this program is the 'file path' statement, which directs shazam to the file that contains the text data set that we are going to access for our analysis ("c:\my2000docs\BYU classes\econ388\classrm_data\utah_cps.asc" is a text file that was previously edited to remove the alphabetic headers over each column, so that it now only contains numeric data). The **file path** statement gives the folder where the data file, utah_cps.asc, resides. Most other new statements just manipulate data into new variables useful for the regression. Male and white variables are dummy variables: they take the value 1 if the condition holds, and 0 otherwise. So that

```
if (sex.eq.1)
```

means that if the condition in the equation is true, then

```
male=1
```

So females (sex=2) have a male=0 value. Race has more than just two possible values; I only created dummy variables for the largest demographic group (but in states with more minorities, you probably would also want to add a dummy variable for blacks, and possibly another one for Hispanics). The "**log(wklywg)**" means that you take the natural logarithm of the wklywg variable, and assign it the name lnwage. The **skipif(wklywg.le.0)** means that if the condition in the parenthesis is true (if wklywg takes a value of zero or less), then this observation will be skipped in the analysis. You have to do that because the logarithms of zero or negative values is not defined.

The Shazam output for the first program is:

```

|_*sample program to illustrate shazam's basic procedures
|_sample 1 20
|_read wage college age male married
   5 VARIABLES AND          20 OBSERVATIONS STARTING AT OBS      1

|_* "stat" does descriptive stats for indicated var
|_stat wage college age male married
NAME      N      MEAN      ST. DEV      VARIANCE      MINIMUM      MAXIMUM
WAGE      20      555.30      280.03      78418.        64.000      971.00
COLLEGE   20      0.55000     0.51042     0.26053       0.0000     1.0000
AGE       20      42.800     15.810     249.96       21.000     73.000
MALE     20      0.45000     0.51042     0.26053       0.0000     1.0000
MARRIED  20      0.60000     0.50262     0.25263       0.0000     1.0000
|_* "ols" does regression with first var=dep. var
|_ols wage college age male

```

REQUIRED MEMORY IS PAR= 3 CURRENT PAR= 1000
 OLS ESTIMATION
 20 OBSERVATIONS DEPENDENT VARIABLE= WAGE
 ...NOTE...SAMPLE RANGE SET TO: 1, 20

R-SQUARE = 0.5726 R-SQUARE ADJUSTED = 0.4924
 VARIANCE OF THE ESTIMATE-SIGMA**2 = 39805.
 STANDARD ERROR OF THE ESTIMATE-SIGMA = 199.51
 SUM OF SQUARED ERRORS-SSE= 0.63688E+06
 MEAN OF DEPENDENT VARIABLE = 555.30
 LOG OF THE LIKELIHOOD FUNCTION = -132.065

VARIABLE NAME	ESTIMATED COEFFICIENT	STANDARD ERROR	T-RATIO	P-VALUE	PARTIAL CORR.	STANDARDIZED COEFFICIENT	ELASTICITY AT MEANS
COLLEGE	201.83	90.32	2.235	0.040	0.488	0.3679	0.1999
AGE	10.599	2.916	3.634	0.002	0.672	0.5984	0.8169
MALE	-85.227	89.70	-0.9501	0.356	-0.231	-0.1553	-0.0691
CONSTANT	28.996	142.7	0.2032	0.842	0.051	0.0000	0.0522

|_* "logit" does qualitative dep var regression

|_logit married college age male

REQUIRED MEMORY IS PAR= 2 CURRENT PAR= 1000
 FOR MAXIMUM EFFICIENCY USE AT LEAST PAR= 3
 LOGIT ANALYSIS DEPENDENT VARIABLE =MARRIED CHOICES = 2
 20. TOTAL OBSERVATIONS
 12. OBSERVATIONS AT ONE
 8. OBSERVATIONS AT ZERO
 25 MAXIMUM ITERATIONS
 CONVERGENCE TOLERANCE =0.00100

LOG OF LIKELIHOOD WITH CONSTANT TERM ONLY = -13.460
 BINOMIAL ESTIMATE = 0.6000
 ITERATION 0 LOG OF LIKELIHOOD FUNCTION = -13.460

ITERATION 1 ESTIMATES
 -0.51502 0.59233E-02 -1.1694 0.96144
 ITERATION 1 LOG OF LIKELIHOOD FUNCTION = -12.455

ITERATION 2 ESTIMATES
 -0.56392 0.61905E-02 -1.2148 1.0377
 ITERATION 2 LOG OF LIKELIHOOD FUNCTION = -12.450

ITERATION 3 ESTIMATES
 -0.56484 0.61930E-02 -1.2159 1.0391

VARIABLE NAME	ESTIMATED COEFFICIENT	ASYMPTOTIC STANDARD ERROR	T-RATIO	ELASTICITY AT MEANS	WEIGHTED AGGREGATE ELASTICITY
COLLEGE	-0.56484	0.98009	-0.57632	-0.12123	-0.11668
AGE	0.61930E-02	0.31264E-01	0.19809	0.10343	0.95481E-01
MALE	-1.2159	0.96360	-1.2618	-0.21350	-0.22023
CONSTANT	1.0391	1.5634	0.66468	0.40549	0.37445

LOG-LIKELIHOOD FUNCTION = -12.450

LOG-LIKELIHOOD(0) = -13.460
 LIKELIHOOD RATIO TEST = 2.02076 WITH 3 D.F.

MADDALA R-SQUARE 0.9610E-01
 CRAGG-UHLER R-SQUARE 0.12991
 MCFADDEN R-SQUARE 0.75064E-01
 ADJUSTED FOR DEGREES OF FREEDOM -0.98361E-01
 APPROXIMATELY F-DISTRIBUTED 0.10821 WITH 3 AND 4 D.F.
 CHOW R-SQUARE 0.96264E-01

PREDICTION SUCCESS TABLE

		ACTUAL	
		0	1
PREDICTED	0	4.	3.
	1	4.	9.

NUMBER OF RIGHT PREDICTIONS = 13.0
 PERCENTAGE OF RIGHT PREDICTIONS = 0.65000

EXPECTED OBSERVATIONS AT 0 = 8.0 OBSERVED = 8.0
 EXPECTED OBSERVATIONS AT 1 = 12.0 OBSERVED = 12.0
 SUM OF SQUARED "RESIDUALS" = 4.3379
 WEIGHTED SUM OF SQUARED "RESIDUALS" = 19.843

HENSHER-JOHNSON PREDICTION SUCCESS TABLE

ACTUAL	PREDICTED	CHOICE		OBSERVED COUNT	OBSERVED SHARE
		0	1		
0		3.669	4.331	8.000	0.400
1		4.331	7.669	12.000	0.600
PREDICTED COUNT		8.000	12.000	20.000	1.000
PREDICTED SHARE		0.400	0.600	1.000	
PROP. SUCCESSFUL		0.459	0.639	0.567	
SUCCESS INDEX		0.059	0.039	0.047	
PROPORTIONAL ERROR		0.000	0.000		
NORMALIZED SUCCESS INDEX				0.098	

|_stop
 TYPE COMMAND

The shazam output from the second program is:

REQUIRED MEMORY IS PAR= 88 CURRENT PAR= 1000
 OLS ESTIMATION
 192 OBSERVATIONS DEPENDENT VARIABLE= LNWAGE
 ...NOTE...SAMPLE RANGE SET TO: 1, 1110

R-SQUARE = 0.1205 R-SQUARE ADJUSTED = 0.1065
 VARIANCE OF THE ESTIMATE-SIGMA**2 = 0.56241
 STANDARD ERROR OF THE ESTIMATE-SIGMA = 0.74994
 SUM OF SQUARED ERRORS-SSE= 105.73
 MEAN OF DEPENDENT VARIABLE = 6.0492
 LOG OF THE LIKELIHOOD FUNCTION = -215.164

MODEL SELECTION TESTS - SEE JUDGE ET AL. (1985,P.242)
 AKAIKE (1969) FINAL PREDICTION ERROR - FPE = 0.57412

(FPE IS ALSO KNOWN AS AMEMIYA PREDICTION CRITERION - PC)
 AKAIKE (1973) INFORMATION CRITERION - LOG AIC = -0.55491
 SCHWARZ (1978) CRITERION - LOG SC = -0.48705
 MODEL SELECTION TESTS - SEE RAMANATHAN (1992,P.167)
 CRAVEN-WAHBA (1979)
 GENERALIZED CROSS VALIDATION - GCV = 0.57437
 HANNAN AND QUINN (1979) CRITERION = 0.59012
 RICE (1984) CRITERION = 0.57463
 SHIBATA (1981) CRITERION = 0.57364
 SCHWARZ (1978) CRITERION - SC = 0.61444
 AKAIKE (1974) INFORMATION CRITERION - AIC = 0.57412

ANALYSIS OF VARIANCE - FROM MEAN				
	SS	DF	MS	F
REGRESSION	14.488	3.	4.8292	8.587
ERROR	105.73	188.	0.56241	P-VALUE
TOTAL	120.22	191.	0.62943	0.000

ANALYSIS OF VARIANCE - FROM ZERO				
	SS	DF	MS	F
REGRESSION	7040.4	4.	1760.1	3129.585
ERROR	105.73	188.	0.56241	P-VALUE
TOTAL	7146.1	192.	37.220	0.000

VARIABLE	ESTIMATED	STANDARD	T-RATIO	PARTIAL STANDARDIZED ELASTICITY			
NAME	COEFFICIENT	ERROR	188 DF	P-VALUE	CORR.	COEFFICIENT	AT MEANS
AGE	0.18501E-01	0.4318E-02	4.285	0.000	0.298	0.2947	0.1134
WHITE	-0.31729	0.3790	-0.8371	0.404	-0.061	-0.0573	-0.0514
MALE	0.33328	0.1093	3.049	0.003	0.217	0.2097	0.0301
CONSTANT	5.4919	0.4117	13.34	0.000	0.697	0.0000	0.9079

|_end
 |_stop
 TYPE COMMAND