

RUNNING AND INTERPRETING WAGE REGRESSIONS USING SHAZAM (using a Utah data set drawn from the 1999 CPS March Supplement)

This discussion assumes that you have read my Shazam_notes.doc, and that you understand what the commands discussed there do. For more information about writing and using Shazam programs, go to:

<http://shazam.econ.ubc.ca>

and click on "Online Examples"; then explore both under "Shazam Guide" (most of Part I, and the Ordinary Least Squares and dummy variables portions of Part II) and under "Examples to Accompany the Manual" (easy to use, arranged by topic).

In these notes, we concentrate on preparing CPS data for manipulation to run $\log(\text{wage})$ and wage regressions, and discuss the interpretation of the resulting regressions. In these wage regressions, there are two types of variables: quantitative and qualitative. Variables that can meaningfully interpreted as having meaningful numeric values like age, tenure, income, weekly wage, number of children under 18 years of age, can be entered into the regression without having to manipulate the variables. These are sometimes called *quantitative* variables. But some variables don't have a natural scale that can be unambiguously matched with the real number line. Gender, occupation, industry, and race are some examples of such variables that have no "natural" numeric values. To measure the impact of such *qualitative* variables—or to simply include them in the analysis to control for their influence—we need to construct **dummy variables**. Dummy variables take a value of "1" if the condition holds, and 0" if it doesn't hold. In the "wage_ex.sha" program below, we create a number of these dummy variables to include them into the analysis. For analyses of microdata such as the CPS, it is not uncommon to have the dummy independent variables outnumber the number of quantitative independent variables.

We illustrate the construction of dummy variables for three groups of qualitative variables: educational attainment (A_HGA), occupation (A_MJOCC) and industry (A_MYIND). The codes for each of these sets of variables is given as follows (and will be employed in the shazam program given below these codes):

A_HGA Demographics, Educational attainment
0 Children
31 Less than 1st grade
32 1st,2nd,3rd,or 4th grade
33 5th or 6th grade
34 7th and 8th grade
35 9th grade
36 10th grade
37 11th grade
38 12th grade no diploma
39 High school graduate-high school diploma
40 Some college but no degree
41 Assc degree-occupation/vocation
42 Assc degree-academic program
43 Bachelor's degree (BA,AB,BS)
44 Master's degree (MA,MS,MENG,MED,MSW,MBA)
45 Professional school degree (MD,DDS,DVM,L
46 Doctorate degree (PHD,EDD)

A_MJOCC Major occupation code
0 Not in univ for children or Armed Forces
1 Executive, Admin, & Managerial Occs
2 Professional Specialty Occs
3 Technicians And Related Support Occs
4 Sales Occs
5 Admin. Support Occs, Incl. Clerical
6 Private Household Occs
7 Protective Service Occs
8 Service Occs, Exc. Protective & Hhld
9 Precision Prod., Craft & Repair Occs
10 Machine Opers, Assemblers & Inspectors
11 Transportation And Material Moving Occs
12 Handlers, equip Cleaners, helpers, laborrs
13 Farming, Forestry And Fishing Occs
14 Armed Forces

A_MJIND Major industry code
0 Not in universe or children
1 Agriculture
2 Mining
3 Construction
4 Manufacturing - Durable Goods
5 Manufacturing- Non-Durable Goods
6 Transportation
7 Communications
8 Utilities And Sanitary Services
9 Wholesale Trade
10 Retail Trade
11 Finance, Insurance, And Real Estate
12 Private Households
13 Business, Auto And Repair Services
14 Personal Services, Exc. Private Hhlds
15 Entertainment And Recreation Services
16 Hospitals
17 Medical Services, Exc. Hospitals
18 Educational Services
19 Social Services
20 Other Professional Services
21 Forestry And Fisheries
22 Public Administration
23 Armed Forces

Suppose that my CPS data was called Utah.txt, and it was on a 3.5" diskette in my a-drive. Further suppose that you want to compare wages by the following educational breakdowns:

No high school degree

High school degree only
Some college but no 4-year college degree
College degree
(the omitted group: those with post-graduate schooling)

Further, suppose you wanted to breakdown occupations into the following groups

Executives
Technical and _sales
Service occupations
Operatives and skilled labor
(the omitted category is mostly unskilled laborers and farmers)

and industries were broken down into the following groups

agriculture, construction, mining and forestry
manufacturing
wholesale and retail trade
public administration
(the omitted industries are mostly service-type industries)

The following shazam program creates these dummy variables (based on the codes provided in the codebook I downloaded when I downloaded the data from the ferret website—see my notes to lecture 1), prints out their descriptive statistics, and then runs a regression with wage as the dependent variable, and a regression with log of wages as the dependent variable. The program follows (with the interpretation of the results discussed below after each regression):

The shazam program:

```
file path a:\
sample 1 1117
read(utah.txt) wklywg mjind mjocc age school marital 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)

* dummy variables for educational attainment follow, based on CPS codes
* the first one indicates to skip those who are children (A_HGA=0)
* the omitted category are those with graduate training (beyond college)
skipif (school.eq.0)
if ((school.eq.31).or.(school.eq.32).or.(school.eq.33).or.(school.eq.34).or.&
(school.eq.35).or.(school.eq.36).or.(school.eq.37).or.(school.eq.38)) no_hi_sc=1
if ((school.ne.31).and.(school.ne.32).and.(school.ne.33).and.(school.ne.34).and.&
(school.ne.35).and.(school.ne.36).and.(school.ne.37).and.(school.ne.38)) no_hi_sc=0
if (school.eq.39) high_sch=1
if (school.ne.39) high_sch=0
if ((school.eq.40).or.(school.eq.41).or.(school.eq.42)) some_col=1
if ((school.ne.40).and.(school.ne.41).and.(school.ne.42)) some_col=0
if (school.eq.43) college=1
if (school.ne.43) college=0
```

```
*the occupational controls follow. these are dummy variables based on A_MJOCC codes
* the first one indicates to skip those who are children or in the armed forces
```

```

* eq means equal; ne means not equal; omitted category is laborers, farmers, foresters
skipif (mjocc.eq.14).or.(mjocc.eq.0)
if ((mjocc.eq.1).or.(mjocc.eq.2)) exec=1
if ((mjocc.ne.1).and.(mjocc.ne.2)) exec=0
if ((mjocc.eq.3).or.(mjocc.eq.4)) tech_sal=1
if ((mjocc.ne.3).and.(mjocc.ne.4)) tech_sal=0
if ((mjocc.eq.5).or.(mjocc.eq.6).or.(mjocc.eq.7).or.(mjocc.eq.8)) serv_occ=1
if ((mjocc.ne.5).and.(mjocc.ne.6).and.(mjocc.ne.7).and.(mjocc.ne.8)) serv_occ=0
if ((mjocc.eq.9).or.(mjocc.eq.10).or.(mjocc.eq.11)) oper_occ=1
if ((mjocc.ne.9).and.(mjocc.ne.10).and.(mjocc.ne.11)) oper_occ=0

```

```

*the industry controls follow. these are dummy variables based on A_MJIND codes
* the first one indicates to skip those who are children or not employed
* eq means equal; ne means not equal; omitted category is laborers, farmers, foresters
skipif (mjind.eq.0)
if ((mjind.eq.1).or.(mjind.eq.2).or.(mjind.eq.3).or.(mjind.eq.21)) ag_cnstr=1
if ((mjind.ne.1).and.(mjind.ne.2).and.(mjind.ne.3).and.(mjind.ne.21)) ag_cnstr=0
if ((mjind.eq.4).or.(mjind.eq.5)) manuf=1
if ((mjind.ne.4).and.(mjind.ne.5)) manuf=0
if ((mjind.eq.9).or.(mjind.eq.10)) trade=1
if ((mjind.ne.9).and.(mjind.ne.10)) trade=0
if (mjind.eq.22) pub_admn=1
if (mjind.ne.22) pub_admn=0

```

```
genr lnwage = log(wklywg)
```

```
stat wklywg lnwage no_hi_sc high_sch some_col college exec tech_sal serv_occ oper_occ &
ag_cnstr manuf trade pub_admn
```

```
ols wklywg age white male no_hi_sc high_sch some_col college exec tech_sal serv_occ &
oper_occ ag_cnstr manuf trade pub_admn
```

```
ols lnwage age white male no_hi_sc high_sch some_col college exec tech_sal serv_occ &
oper_occ ag_cnstr manuf trade pub_admn
```

```
end
stop
```

The shazam output:

```

***some welcoming/data set stuff precedes, and then ****
|_stat wklywg lnwage no_hi_sc high_sch some_col college exec tech_sal serv_occ oper_occ &
|_ag_cnstr manuf trade pub_admn
NAME          N      MEAN      ST. DEV      VARIANCE      MINIMUM      MAXIMUM
WKLYWG        194    543.14      369.23      0.13633E+06    10.000      2000.0
LNWAGE        194     6.0482      0.78993     0.62399        2.3026      7.6009
NO_HI_SC      924    0.93074E-01 0.29069     0.84502E-01    0.0000      1.0000
HIGH_SCH      924    0.29004     0.45403     0.20614        0.0000      1.0000
SOME_COL      924    0.35823     0.47974     0.23015        0.0000      1.0000
COLLEGE       924    0.17749     0.38229     0.14614        0.0000      1.0000
EXEC           924    0.31494     0.46474     0.21598        0.0000      1.0000
TECH_SAL      924    0.15368     0.36084     0.13020        0.0000      1.0000
SERV_OCC      924    0.26082     0.43932     0.19300        0.0000      1.0000
OPER_OCC      924    0.22186     0.41572     0.17283        0.0000      1.0000
AG_CNSTR      924    0.11364     0.31754     0.10083        0.0000      1.0000
MANUF         924    0.13636     0.34336     0.11790        0.0000      1.0000
TRADE         924    0.21320     0.40979     0.16793        0.0000      1.0000

```

```
PUB_ADMN      924  0.48701E-01  0.21536      0.46380E-01  0.0000      1.0000
```

Discussion: the stuff above represents the descriptive statistics for Utah, when the population is restricted to those between 18 and 65 years of age. It is good to check your data to make sure the mean and range (minimum, maximums) are what you expect. Sometimes errors are made—best to catch them before worrying about interpreting the regression results. Note that the dummy variables all have values between 0 and 1, and their mean is the fraction of the population with that characteristic: 29 percent have a high school diploma and no further education; 31.5 percent are in executive occupations.

Now for the first regression:

```
|_ols wklywg age white male no_hi_sc high_sch some_col college exec tech_sal serv_occ
oper_occ &
```

```
| ag_cnstr manuf trade pub_admn
REQUIRED MEMORY IS PAR=      248 CURRENT PAR=      1000
```

```
OLS ESTIMATION
```

```
194 OBSERVATIONS      DEPENDENT VARIABLE= WKLYWG
```

```
...NOTE...SAMPLE RANGE SET TO:      1, 1117
```

```
R-SQUARE =      0.3320      R-SQUARE ADJUSTED =      0.2757
VARIANCE OF THE ESTIMATE-SIGMA**2 =      98738.
STANDARD ERROR OF THE ESTIMATE-SIGMA =      314.23
SUM OF SQUARED ERRORS-SSE=      0.17575E+08
MEAN OF DEPENDENT VARIABLE =      543.14
LOG OF THE LIKELIHOOD FUNCTION = -1382.45
```

MODEL SELECTION TESTS - SEE JUDGE ET AL. (1985,P.242)

```
AKAIKE (1969) FINAL PREDICTION ERROR - FPE =      0.10688E+06
(FPE IS ALSO KNOWN AS AMEMIYA PREDICTION CRITERION - PC)
```

```
AKAIKE (1973) INFORMATION CRITERION - LOG AIC =      11.579
```

```
SCHWARZ (1978) CRITERION - LOG SC =      11.849
```

MODEL SELECTION TESTS - SEE RAMANATHAN (1992,P.167)

```
CRAVEN-WAHBA (1979)
```

```
GENERALIZED CROSS VALIDATION - GCV =      0.10761E+06
```

```
HANNAN AND QUINN (1979) CRITERION =      0.11916E+06
```

```
RICE (1984) CRITERION =      0.10849E+06
```

```
SHIBATA (1981) CRITERION =      0.10554E+06
```

```
SCHWARZ (1978) CRITERION - SC =      0.13989E+06
```

```
AKAIKE (1974) INFORMATION CRITERION - AIC =      0.10684E+06
```

ANALYSIS OF VARIANCE - FROM MEAN

	SS	DF	MS	F
REGRESSION	0.87359E+07	15.	0.58239E+06	5.898
ERROR	0.17575E+08	178.	98738.	P-VALUE
TOTAL	0.26311E+08	193.	0.13633E+06	0.000

ANALYSIS OF VARIANCE - FROM ZERO

	SS	DF	MS	F
REGRESSION	0.65967E+08	16.	0.41229E+07	41.756
ERROR	0.17575E+08	178.	98738.	P-VALUE
TOTAL	0.83542E+08	194.	0.43063E+06	0.000

VARIABLE NAME	ESTIMATED COEFFICIENT	STANDARD ERROR	T-RATIO	PARTIAL CORR. COEFFICIENT	STANDARDIZED ELASTICITY AT MEANS
AGE	5.5847	1.928	2.896	0.004 0.212	0.1949 0.3841

WHITE	-114.28	161.1	-0.7095	0.479-0.053	-0.0441	-0.2061
MALE	139.93	50.82	2.754	0.007 0.202	0.1890	0.1421
NO_HI_SC	-460.79	127.9	-3.602	0.000-0.261	-0.3442	-0.0700
HIGH_SCH	-416.09	102.7	-4.053	0.000-0.291	-0.5064	-0.2132
SOME_COL	-328.46	96.49	-3.404	0.001-0.247	-0.4373	-0.2431
COLLEGE	-172.81	101.6	-1.700	0.091-0.126	-0.1719	-0.0508
EXEC	147.55	135.5	1.089	0.278 0.081	0.1918	0.0966
TECH_SAL	-21.972	141.5	-0.1552	0.877-0.012	-0.0189	-0.0046
SERV_OCC	-24.277	135.3	-0.1794	0.858-0.013	-0.0282	-0.0108
OPER_OCC	47.833	129.8	0.3684	0.713 0.028	0.0560	0.0218
AG_CNSTR	110.54	88.97	1.242	0.216 0.093	0.0892	0.0199
MANUF	56.721	71.81	0.7898	0.431 0.059	0.0549	0.0156
TRADE	99.661	72.15	1.381	0.169 0.103	0.1063	0.0350
PUB_ADMN	79.305	102.9	0.7709	0.442 0.058	0.0498	0.0083
CONSTANT	584.02	242.2	2.411	0.017 0.178	0.0000	1.0753

Discussion of the first (wage) regression results: The R-square indicates that 33.2 percent of the variation in wages (across individuals) can be explained by the variation in the dependent variables. The adjusted R-square is used to compare models using the same dependent variable (here, wages—not log wages) with the same sample size. In the “Analysis of Variance – From Mean” section, latter on in this course we will learn how to use the sum of squared errors (SSE=17,575,000) and degrees of freedom (DF=178) to do tests of joint significance of several independent variables simultaneously.

But most important are the estimated coefficients and their associated t-statistics (and p-values). The estimated coefficients are indicate how much the wage changes with a unit increase (1-point) in the associated independent variables, holding the values of the other variables constant. So, for example:

- 5.5857=for each additional year of AGE, the usual weekly wage increases by \$5.59
- 114.28=whites make \$114.28 less per week on average than nonwhites (not statistically significant)
- 139.93=males make \$139.93 more than females (the omitted group) per week on average
- 460.79=non-high school grads make \$460.79 less per week than those with post graduate work
- 172.81=college graduates make \$172.81 less per week than those with post-grad work (omitted grp)
- 147.55=executives make \$147.55 more per week than unskilled labor and farmers (the omitted group)
- 47.83=skilled occupations make \$47.83 more than unskilled labor and farmers (the omitted group)
- 99.66=those in wholesale or retail trade industries make \$99.61 more than those in the service industries

T-values approximately greater than 2 or smaller than minus 2 (indicating more than two standard deviations away from a coefficient value of zero) are statistically significant at the 5 percent level. They will also have p-values of .050000 or smaller. P-value is the probability of making a type I error, or the likelihood of getting the sample result that we did if the null hypothesis were true. The statistically significant independent variables at the 5-percent level are: age, male, no_hi_sc (no high school diploma), high-sch (high school diploma only), some_col (those with some college but not a 4-year degree) if they are significant at the 5-percent level. Their t-values also indicate statistical significance for these variables at the one percent level. College is significant at the 10 percent level of signficiance, but not the 5 percent level.

Now for the second (log(wage)) regression:

```
|_ols lnwage age white male no_hi_sc high_sch some_col college exec tech_sal serv_occ
oper_occ &
| ag_cnstr manuf trade pub_admn
```

```
REQUIRED MEMORY IS PAR=      248 CURRENT PAR=      1000
```

OLS ESTIMATION

194 OBSERVATIONS DEPENDENT VARIABLE= LNWAGE

...NOTE..SAMPLE RANGE SET TO: 1, 1117

R-SQUARE = 0.2658 R-SQUARE ADJUSTED = 0.2039
 VARIANCE OF THE ESTIMATE-SIGMA**2 = 0.49677
 STANDARD ERROR OF THE ESTIMATE-SIGMA = 0.70482
 SUM OF SQUARED ERRORS-SSE= 88.425
 MEAN OF DEPENDENT VARIABLE = 6.0482
 LOG OF THE LIKELIHOOD FUNCTION = -199.061

MODEL SELECTION TESTS - SEE JUDGE ET AL. (1985,P.242)

AKAIKE (1969) FINAL PREDICTION ERROR - FPE = 0.53774
 (FPE IS ALSO KNOWN AS AMEMIYA PREDICTION CRITERION - PC)
 AKAIKE (1973) INFORMATION CRITERION - LOG AIC = -0.62075
 SCHWARZ (1978) CRITERION - LOG SC = -0.35124

MODEL SELECTION TESTS - SEE RAMANATHAN (1992,P.167)

CRAVEN-WAHBA (1979)
 GENERALIZED CROSS VALIDATION - GCV = 0.54142
 HANNAN AND QUINN (1979) CRITERION = 0.59952
 RICE (1984) CRITERION = 0.54583
 SHIBATA (1981) CRITERION = 0.53098
 SCHWARZ (1978) CRITERION - SC = 0.70382
 AKAIKE (1974) INFORMATION CRITERION - AIC = 0.53754

ANALYSIS OF VARIANCE - FROM MEAN

	SS	DF	MS	F
REGRESSION	32.005	15.	2.1337	4.295
ERROR	88.425	178.	0.49677	P-VALUE
TOTAL	120.43	193.	0.62399	0.000

ANALYSIS OF VARIANCE - FROM ZERO

	SS	DF	MS	F
REGRESSION	7128.6	16.	445.54	896.867
ERROR	88.425	178.	0.49677	P-VALUE
TOTAL	7217.0	194.	37.201	0.000

VARIABLE	ESTIMATED	STANDARD	T-RATIO	PARTIAL		STANDARDIZED	ELASTICITY
NAME	COEFFICIENT	ERROR	178 DF	P-VALUE	CORR.	COEFFICIENT	AT MEANS
AGE	0.10519E-01	0.4325E-02	2.432	0.016	0.179	0.1716	0.0650
WHITE	-0.36870	0.3613	-1.021	0.309	-0.076	-0.0665	-0.0597
MALE	0.21933	0.1140	1.924	0.056	0.143	0.1384	0.0200
NO_HI_SC	-0.81346	0.2870	-2.835	0.005	-0.208	-0.2840	-0.0111
HIGH_SCH	-0.79695	0.2303	-3.461	0.001	-0.251	-0.4533	-0.0367
SOME_COL	-0.57350	0.2164	-2.650	0.009	-0.195	-0.3569	-0.0381
COLLEGE	-0.36988	0.2280	-1.622	0.107	-0.121	-0.1720	-0.0098
EXEC	0.27561	0.3040	0.9065	0.366	0.068	0.1675	0.0162
TECH_SAL	-0.88560E-01	0.3175	-0.2789	0.781	-0.021	-0.0356	-0.0017
SERV_OCC	-0.13590	0.3035	-0.4478	0.655	-0.034	-0.0739	-0.0054
OPER_OCC	0.13332	0.2912	0.4578	0.648	0.034	0.0730	0.0055
AG_CNSTR	0.39859	0.1996	1.997	0.047	0.148	0.1504	0.0065
MANUF	0.19594	0.1611	1.216	0.225	0.091	0.0887	0.0048
TRADE	0.29618	0.1618	1.830	0.069	0.136	0.1477	0.0093
PUB_ADMN	0.26805	0.2308	1.162	0.247	0.087	0.0787	0.0025
CONSTANT	6.2459	0.5433	11.50	0.000	0.653	0.0000	1.0327

|_end

l_stop

Discussion of the second (log wage) regression results: The R-square indicates that 26.58 percent of the variation in wages (across individuals) can be explained by the variation in the dependent variables. In the “Analysis of Variance – From Mean” section, useful to do tests of joint significance of several independent variables simultaneously, the sum of squared errors (SSE) equals 88.425 and degrees of freedom (DF) equals 178.

A unit change in a logarithm is the percentage change in that variable. This changes the interpretation of the coefficients slightly when we have $\log(\text{wages})$ as the dependent variable: the estimated coefficients indicate the percentage change in wages given a unit increase (1-point) in the associated independent variables, holding the values of the other variables constant. So, for example:

- .010519=for each additional year of AGE, the usual weekly wage increases by 1.05 percent
- .3687=whites make 36.87 percent less per week on average than nonwhites (not statistically significant)
- .21933=males make 21.9 percent more than females (the omitted group) per week on average
- .81346=non-high school grads make 81.3 percent less per week than those with post graduate work
- .36988=college graduates 37 percent less per week than those with post-grad work (omitted group)
- .27561=executives make 27.6 percent more per week than unskilled labor and farmers (omitted group)
- .1333=skilled occupations make 13.3 percent more than unskilled labor and farmers (the omitted group)
- .296=those in trade industries make 29.6 percent more than those in the service industries (omitted grp)

T-values approximately greater than 2 or smaller than minus 2 (indicating more than two standard deviations away from a coefficient value of zero) are statistically significant at the 5 percent level, as they do in any regression. They will also have p-values of .050000 or smaller if they are significant at the 5-percent level. The statistically significant independent variables at the 5-percent level are: age, no_hi_sc (no high school diploma), high_sch (high school diploma only), some_col (those with some college but not a 4-year degree), and the ag_constr (those in agricultural or construction industries). Male and trade industry dummy variables are significant at the 10 percent level of significance, but not the 5 percent level.