8.9 Exercises

1. Use gss2006 chapter8.dta. Imagine that you heard somebody say that there was no reason to provide more educational opportunities for women because so many of them just stay at home anyway. You have a variable measuring education, educ, and a variable measuring hours worked in the last week, hrs1. Do a correlation and regression of hours worked in the last week on years of education. Then do this separately for women and for men. Interpret the correlation and the slope for the overall sample and then for women and for men separately. Is there an element of truth to what you heard?

Mean Distribution of Data

Distribution of worked hours



Distribution of worked hours for Males



Distribution of hours worked for Females



Distribution of Education



Distribution of Education for Males



Distribution of Education for Females



Structure of the Data for Analysis to see if there are any missing values or special values

. codebook hrs1

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

hrs1 NUMBER OF HOURS WORKED LAST WEEK

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

type: numeric (byte)

label: HRS1, but 88 nonmissing values are not labeled

range: [1,89] units: 1

unique values: 88 missing .: 3,776/10,179

examples: 40

45

70

.

. codebook educ

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

educ HIGHEST YEAR OF SCHOOL COMPLETED

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

type: numeric (byte)

label: EDUC, but 21 nonmissing values are not labeled

range: [0,20] units: 1

unique values: 21 missing .: 32/10,179

examples: 12

12

14

16

. codebook sex

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

sex RESPONDENTS SEX

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

type: numeric (byte)

label: SEX

range: [1,2] units: 1

unique values: 2 missing .: 0/10,179

tabulation: Freq. Numeric Label

4,516 1 MALE

5,663 2 FEMALE

.

Conclusion of the Data Analysis:There are missing values in both the variables

Correlation test for both gendas

Males correlation test

. pwcorr hrs1 educ if sex==1

| hrs1 educ

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

hrs1 | 1.0000

educ | 0.0681 1.0000

Females correlation test

. pwcorr hrs1 educ if sex==2

| hrs1 educ

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

hrs1 | 1.0000

educ | 0.1220 1.0000

Regression Test

. regress educ hrs1, beta

Source | SS df MS Number of obs = 6,387

-------------+---------------------------------- F(1, 6385) = 44.86

Model | 374.7339 1 374.7339 Prob > F = 0.0000

Residual | 53335.8949 6,385 8.35331165 R-squared = 0.0070

-------------+---------------------------------- Adj R-squared = 0.0068

Total | 53710.6288 6,386 8.41068412 Root MSE = 2.8902

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

educ | Coef. Std. Err. t P>|t| Beta

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

hrs1 | .0169341 .0025283 6.70 0.000 .0835279

\_cons | 13.06539 .1123928 116.25 0.000 .

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

Regression Test for Males

. regress educ hrs1 if sex==1, beta

Source | SS df MS Number of obs = 3,195

-------------+---------------------------------- F(1, 3193) = 14.86

Model | 138.959577 1 138.959577 Prob > F = 0.0001

Residual | 29863.1706 3,193 9.35269985 R-squared = 0.0046

-------------+---------------------------------- Adj R-squared = 0.0043

Total | 30002.1302 3,194 9.39327808 Root MSE = 3.0582

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

educ | Coef. Std. Err. t P>|t| Beta

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

hrs1 | .0145222 .0037675 3.85 0.000 .0680563

\_cons | 13.0301 .1785737 72.97 0.000 .

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

.

Regression test for females

. regress educ hrs1 if sex==2, beta

Source | SS df MS Number of obs = 3,192

-------------+---------------------------------- F(1, 3190) = 48.17

Model | 351.844687 1 351.844687 Prob > F = 0.0000

Residual | 23302.4598 3,190 7.30484634 R-squared = 0.0149

-------------+---------------------------------- Adj R-squared = 0.0146

Total | 23654.3045 3,191 7.41281871 Root MSE = 2.7027

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

educ | Coef. Std. Err. t P>|t| Beta

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

hrs1 | .0244788 .0035271 6.94 0.000 .1219608

\_cons | 12.91548 .1456582 88.67 0.000 .

Summary Analytics : Females have a higher beta value and correlation-coefficient than males therefore the assumptions that there is no need to educate them a they will just stay at home is wrong.

.

1. Use gss2006 chapter8.dta. What is the relationship between the hours a person works and the hours his or her spouse works? Do this for women and for men separately. Compute the correlation, the regression results, and the scattergrams. Interpret each of these. Next test if the correlation is statistically significant and interpret the results.

Structure of the Data

. codebook sphrs1

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

sphrs1 NUMBER OF HRS SPOUSE WORKED LAST WEEK

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

type: numeric (byte)

label: SPHRS1, but 78 nonmissing values are not labeled

range: [1,89] units: 1

unique values: 78 missing .: 6,961/10,179

examples: 40

.

.

.

. codebook hrs1

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

hrs1 NUMBER OF HOURS WORKED LAST WEEK

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

type: numeric (byte)

label: HRS1, but 88 nonmissing values are not labeled

range: [1,89] units: 1

unique values: 88 missing .: 3,776/10,179

examples: 40

45

70

.

Data Structure Analysis :There are missing values between the two variables

Scatter-grams of the two variables



Scatter-gram for the two variable for Males



Scatter-gram of the two variable for Females



Regression Analysis

. regress sphrs1 hrs1, beta

Source | SS df MS Number of obs = 2,393

-------------+---------------------------------- F(1, 2391) = 0.75

Model | 130.596748 1 130.596748 Prob > F = 0.3872

Residual | 417393.396 2,391 174.568547 R-squared = 0.0003

-------------+---------------------------------- Adj R-squared = -0.0001

Total | 417523.992 2,392 174.550164 Root MSE = 13.212

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

sphrs1 | Coef. Std. Err. t P>|t| Beta

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

hrs1 | .0164893 .0190642 0.86 0.387 .0176858

\_cons | 40.7701 .8522983 47.84 0.000 .

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

Regression Analysis for Males

Regression . regress sphrs1 hrs1 if sex==1, beta

Source | SS df MS Number of obs = 1,136

-------------+---------------------------------- F(1, 1134) = 10.86

Model | 1752.10969 1 1752.10969 Prob > F = 0.0010

Residual | 182950.073 1,134 161.331634 R-squared = 0.0095

-------------+---------------------------------- Adj R-squared = 0.0086

Total | 184702.183 1,135 162.733201 Root MSE = 12.702

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

sphrs1 | Coef. Std. Err. t P>|t| Beta

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

hrs1 | .0991445 .0300848 3.30 0.001 .0973968

\_cons | 31.98229 1.472278 21.72 0.000 .

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

Analysis for Females

. regress sphrs1 hrs1 if sex==2, beta

Source | SS df MS Number of obs = 1,257

-------------+---------------------------------- F(1, 1255) = 48.16

Model | 6765.1294 1 6765.1294 Prob > F = 0.0000

Residual | 176296.727 1,255 140.47548 R-squared = 0.0370

-------------+---------------------------------- Adj R-squared = 0.0362

Total | 183061.857 1,256 145.749886 Root MSE = 11.852

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

sphrs1 | Coef. Std. Err. t P>|t| Beta

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

hrs1 | .1644874 .0237025 6.94 0.000 .1922379

\_cons | 39.55875 .9600615 41.20 0.000 .

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

Correlation Analysis

pwcorr sphrs1 hrs1

| sphrs1 hrs1

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

sphrs1 | 1.0000

hrs1 | 0.0177 1.0000

Correlation Analysis for Males

. pwcorr sphrs1 hrs1 if sex==1

| sphrs1 hrs1

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

sphrs1 | 1.0000

hrs1 | 0.0974 1.0000

Correlation Analysis for Females

. pwcorr sphrs1 hrs1 if sex==2

| sphrs1 hrs1

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

sphrs1 | 1.0000

hrs1 | 0.1922 1.0000

Summary Analytics:

1. Use gss2006 chapter8.dta. Repeat figure 8.2 using your own subsample of 250 observations. Then repeat the figure using a jitter(3) option. Compare the two figures. Set your seed at 111.

Data Structure

. set seed 111

. sample 250 ,count

(9,929 observations deleted)

twoway (scatter educ paeduc if sex==1, msize(5)) if sex==1, ytitle(Son's Education) xtitle(Fa

> thers's Education) title(Scattergram relating father’s education to his son’s education) note

> (N=250 observation)



. twoway (scatter educ paeduc if sex==1, msize(5) jitter(3) jitterseed(111)) if sex==1, ytitle(

> Son's Education) xtitle(Fathers's Education) title(Scattergram relating father’s education to

> his son’s education) note(N=250 observation)



Summary Analytics:

1. . Use gss2006 chapter8.dta. Compute the correlations between happy, hapmar, and health by using correlate and then again by using pwcorr. Why are the results slightly different? Then estimate the correlations by using pwcorr, and get the significance level and the number of observations for each case. Finally, repeat the pwcorr command so that all the Ns are the same (that is, there is casewise/listwise deletion).

Structure of Data

. codebook happy hapmar health

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

happy GENERAL HAPPINESS

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

type: numeric (byte)

label: HAPPY

range: [1,3] units: 1

unique values: 3 missing .: 70/250

tabulation: Freq. Numeric Label

47 1 VERY HAPPY

111 2 PRETTY HAPPY

22 3 NOT TOO HAPPY

70 .

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

hapmar HAPPINESS OF MARRIAGE

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

type: numeric (byte)

label: HAPMAR

range: [1,3] units: 1

unique values: 3 missing .: 168/250

tabulation: Freq. Numeric Label

56 1 VERY HAPPY

24 2 PRETTY HAPPY

2 3 NOT TOO HAPPY

168 .

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

health CONDITION OF HEALTH

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

type: numeric (byte)

label: HEALTH

range: [1,4] units: 1

unique values: 4 missing .: 61/250

tabulation: Freq. Numeric Label

40 1 EXCELLENT

108 2 GOOD

34 3 FAIR

7 4 POOR

61 .

.

Correlations between happy, hapmar, and health by using correlate

. correlate happy hapmar health, means

(obs=53)

Variable | Mean Std. Dev. Min Max

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

happy | 1.528302 .5407891 1 3

hapmar | 1.264151 .4450991 1 2

health | 2 .7337994 1 4

| happy hapmar health

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

happy | 1.0000

hapmar | 0.5276 1.0000

health | 0.0485 0.0000 1.0000

Correlations between happy, hapmar, and health by using pwcorr

. pwcorr happy hapmar health

| happy hapmar health

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

happy | 1.0000

hapmar | 0.5821 1.0000

health | 0.1881 0.0000 1.0000

Estimate the correlations by using pwcorr, and get the significance level and the number of observations for each case

. pwcorr happy hapmar health, obs sig

| happy hapmar health

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

happy | 1.0000

|

| 180

|

hapmar | 0.5821 1.0000

| 0.0000

| 82 82

|

health | 0.1881 0.0000 1.0000

| 0.0349 1.0000

| 126 53 189

|

.

Repeat the pwcorr command so that all the Ns are the same (that is, there is casewise/listwise deletion).

. pwcorr happy hapmar health, obs sig listwise

| happy hapmar health

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

happy | 1.0000

|

| 53

|

hapmar | 0.5276 1.0000

| 0.0000

| 53 53

|

health | 0.0485 0.0000 1.0000

| 0.7304 1.0000

| 53 53 53

|

Summary Analytics:

1. . Use gss2002 chapter8.dta. There are two variables called happy7 and satfam7. Run the codebook command on these variables. Notice how the higher score goes with being unhappy or being dissatisfied. You always want the higher score to mean more of a variable, so generate new variables (happynew and satfamnew) that reverse these codes so that a score of 1 on happynew means very unhappy and a score of 7 means very happy. Similarly, a score of 1 on satfamnew means very dissatisfied and a score of 7 means very satisfied. Now do a regression of happiness on family satisfaction with the new variables. How correlated are these variables? Write the regression equation. Interpret the constant and the slope.
2. Skip
3. Use depression.dta from the Stata Press website; that is, type . use http://www.stata-press.com/data/r13/depression.dta From this hypothetical data, you are interested in the relationship between depression (variable TotalScore) and age. (This dataset uses capitalization as an aid in reading the total score variable. This is rarely a good idea because it is hard to remember these conventions, and if you always use all lowercase, you do not need to remember when and how you used capitalization. Perhaps better options would be to label the variable totalscore or total score.) Are older people more or less depressed? a. Type scatter and binscatter to describe the relationship. b. Interpret these results. Why is the binscatter graph easier to interpret?
4. You suspect that the relationship may be nonlinear with a gradual increase among those over about 50 years of age. a. Type binscatter to fit a curve. b. Interpret these results and compare them with the graphs created in exercise 7.