

May 20 2005

[I]. Levels of analysis: univariate (one variable)→bivariate (2 variables; 2 xs or x & Y)→multivariate (many explanatory variables (e.g., $y=x_1 + x_2 + x_n \dots$)). Pyramid.

[II]. Univariate analysis and how to present results of univariate analysis

- Type of inspection in univariate and bivariate analysis
 - + Visual (graph)
 - scatter wagetot saletot
 - + Descriptive statistics (e.g., STATA: summarize x; summarize x, detail; inspect x)
 - + Simple bivariate statistics (e.g., STATA: tab male [to see shares of females and males])
- Typology of variables:
 - Text: variable was coded with words (e.g., name of person) In STATA “browse lastnam1” – person’s patronym)
 - Categorical (make sure the +1 category=name of variable; e.g., male=1 if person is male; male=0 if person is female. If labeled in this way then when you use the “sum male” the mean value can be interpreted as the share of the population that is male). Other example of categorical variables are political parties, religion, national/racial/ethnic identity
 - Continuous
- Cleaning procedures: coding errors & missing values (costs of missing values)
 - Reduce survey-code book-software for inputing-software for analysis
 - Pre-code: subject identification numbers and acceptable values
- Description of variable
 - Stata commands:
 - tabulate y x, column row chi2 (make sure you understand the chi2 command and what the chi2 distribution stands for)
 - sum x, detail
 - inspect x
 - Mean: for continuous and for binary variables
 - Median: does distribution violate assumption of normality? If so, median might make summary stats robust to outliers
 - Minimum and maximum values: are they reasonable or do they pick up coding errors or true outliers?
 - Standard deviation: spread around the mean. With little variance in x or y, you won’t have much statistical power. Coefficient of variation = $sd/mean$
 - Are variables censored?
 - Left censored variables: how much did you spend buying a car last year? Most people won’t have spend any money, so many zeros
 - Right censored: how many hours did you work last week? Most people who are fully employed will respond 40 hours
 - Double censoring: share of hours spend in house work. Full-time house wives/husbands might spend 100% and at the other extreme some people might spend no time.
 - Are variables truncated?
 - Definition: you don’t see the variable (in censored variables you see the variables, but most of them are zeros)
 - Example: you examine the relation between income and schooling in a village. People with a lot of schooling who earn a lot of income leave the village, so you don’t see the upper end of the schooling or income variables and might conclude that schooling and income are poorly correlated when in fact they are correlated, you just were not able to see the part of the distributions where they were correlated.
 - With truncated variables we need to start thinking about identification instruments – variables that might predict when/why people leave the sample. This is an advanced topic, but worth remembering from the outset.
- Power transformations to correct for positive or negatively skewed distributions
- Univariate analysis should culminate in a table of summary statistics (see Table 1) where you define the variables and provide summary statistics, typically: # of observations/variable, mean, standard deviation, and/or min and max values
- Stata command to obtain descriptive statistics (make sure to include both dependent and explanatory variables)
 - describe y x1 x2 x3
 - sum y x1 x2 x3

[III]. Basic statistical concepts you should know.

- Central limit theorem (as sample size increases, variance/standard deviation shrinks)
- Normality: t, z, and chi2 distribution and their relation to probability
- Null hypothesis and rejection of null

- R2, residual sum of squares, total sum of squares (Some of these are explained in the handout on how to read STATA output)
- Degrees of freedom (# of effective observations you have to estimate a parameter. Every coefficient uses one degree of freedom. E.g., you need two observations to estimate ONE mean value. So the degree of freedoms in a regression = n-k where k is # of variables you want to estimate).

[IV]. Bivariate analysis

- Exploration versus data mining
- End of bivariate regression is to narrow the Ys and Xs. You want parsimony on the right hand side because the inclusion of too many Xs on the right-side will make the statistical results weaker.
- Types:
 - within dependent variable: robustness and consistency
 - within explanatory variable: multicollinearity
 - Defined: excessive overlap between right-side variables
 - We care about it because it affects the estimate of the xs we care about
 - Why do we leave it in, even if it is a problem? Because sometimes we care about the joint effect of group of variables (e.g., we care about how different dimensions of culture – language, values, behaviors – as a group affect a welfare outcome)
 - How to overcome it? (a) principal component analysis – we create one variable (e.g., an index) out of many right-hand side variables, (b) we include them all and then test for joint significance, (c) we use backward or forward process of elimination
 - With any data set, there is an alpha probability of 5% of finding significant results.
 - If you have, say 20 variables, you have $20^2-1=399$ combinations
 - With little time and no theory, people often resort to stepwise regression
 - Stepwise regression
 - between dependent and explanatory variable: eliminate redundant explanatory variables
 - A first test to see if the two variables are related
 - Outliers can cause heteroskedasticity
- A draft typology of regression types/statistical analysis [see [V] below for reading computer output for cells [A]-[D]]:

		Explanatory variables is:	
		Categorical	Continuous
Dependent variable is:	Categorical	Chi2 [A]	Logits, probits [C]
	Continuous	T test, graph [B]	Regression, graph [D]

In Stata (see below):

[A]. `tab y x, column row chi2`

[B]. `ttest height, by(male)` (where 1 if person is men and 0=women). Here you are testing whether the difference in the mean of two populations is statistically significant.

[C]. Save for later; for now treat as in B

[D]. `regress y x` (OLS, tobits, quantile/median regressions)

- A useful first step is to simply correlate all the variables you are thinking of using to see which among them are weakly or highly correlated
 - `pwcorr y1 y2 x1 x2 x3, sig sidak` (The second line gives you the probability distribution. Sig sidak adjusts for multiple comparisons)

```
pwcorr stature weight age, sig sidak
      |  stature  weight  age
-----+-----
stature |  1.0000
weight  |  0.9421  1.0000
      |  0.0000
age     |  0.7038  0.7531  1.0000
      |  0.0000  0.0000
```

[V]. Reading computer output for cells [A]-[D] of bivariate analysis

A. REGRESSION: WHERE BOTH DEPT AND EXP VARIABLES ARE CONTINUOUS [Cell [D]]. Regression of cash income (dept variable) against the average female adult age of the household

$$Y = a + b*x$$

```
. reg cashincometotal agefemavgadulths [STATA COMMAND FOR SIMPLE, OLS REGRESSION]
```

Source	SS	df	MS	Number of obs =	489
Model	177326.98	1	177326.98	F(1, 487) =	3.46
Residual	24972669.0	487	51278.5812	Prob > F =	0.0635
				R-squared =	0.0071

```

-----+-----
Total | 25149996.0  488  51536.8771          Adj R-squared = 0.0050
                                         Root MSE   = 226.45
-----+-----
cashincome~1 |      Coef.   Std. Err.      t    P>|t|    [95% Conf. Interval]
-----+-----
agefemavga~s |   1.516353   .8154193    1.86  0.064   - .0858209   3.118527
  _cons |   94.2613  28.65929    3.29  0.001   37.95018   150.5724
-----+-----

```

NOTE:

- Intercept or constant (94.2)
- Coefficient:
 - o Sign of coefficient (+)
 - o Size of coefficient (1.51)
 - o Reading of coefficient: an additional year of age correlates with 1.51 more bolivianos in earnings
- Sample size(489)
- T statistics (1.86)=coefficient/standard error=1.51/0.815=1.86
- P value (0.064)

[B]. REGRESSION WHERE DEPENDENT VARIABLE (cash income) IS CONTINUOUS BUT EXPLANATORY VARIABLES (CLOSE) IS CATEGORICAL (close=1 if village<25km; close=0 if village>25km)[This corresponds to cell [B]]

$Y = a + b \cdot \text{close}$

```
. reg cashincometotal close
```

```

-----+-----
Source |      SS      df      MS              Number of obs =    511
-----+-----
Model | 189932.829    1 189932.829          F( 1, 509) =    3.80
Residual | 25417784.9  509 49936.7091          Prob > F      = 0.0517
-----+-----
Total | 25607717.7  510 50211.2113          R-squared     = 0.0074
                                         Adj R-squared = 0.0055
                                         Root MSE    = 223.47
-----+-----
cashincome~1 |      Coef.   Std. Err.      t    P>|t|    [95% Conf. Interval]
-----+-----
close |   38.81806  19.90416    1.95  0.052   - .2863508   77.92247
  _cons |   124.6607  13.23694    9.42  0.000   98.65494   150.6665
-----+-----

```

$Y = 124 + 38.8 \cdot \text{close}$

NOTE:

- Represent equation visually
- Estimate Y if close=0; if close=1

Other ways of obtaining the same results:

[I]. Simple summary statistics of income for close and far

```
. sum cashincometotal if close==1
```

```

-----+-----
Variable |      Obs      Mean   Std. Dev.      Min      Max
-----+-----
cashincome~1 |    226   163.4788  225.8498         0    1515
-----+-----

```

```
. sum cashincometotal if close==0
```

```

-----+-----
Variable |      Obs      Mean   Std. Dev.      Min      Max
-----+-----
cashincome~1 |    285   124.6607  221.5578         0    2200
-----+-----

```

[II]. T test to see if the mean difference between the two samples differs in a statistically significant way:

```
. ttest cashincometotal, by(close)
```

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
0	285	124.6607	13.12395	221.5578	98.82814	150.4933
1	226	163.4788	15.0233	225.8498	133.8744	193.0831
combined	511	141.8288	9.912653	224.0786	122.3541	161.3034
diff		-38.81806	19.90416		-77.92247	.2863508

Degrees of freedom: 509

Ho: mean(0) - mean(1) = diff = 0
 Ha: diff < 0 Ha: diff ~= 0 Ha: diff > 0
 t = -1.9502 t = **-1.9502** t = -1.9502
 P < t = 0.0258 P > |t| = **0.0517** P > t = 0.9742

[C]. REGRESSION WHERE DEPENDENT VARIABLE IS CATEGORICAL (HOUSEHOLD ENGAGED IN SOME BARTER/SWAP=1; 0=HOUSEHOLD DID NOT ENGAGE IN ANY BARTER/SWAP) AND EXPLANATORY VARIABLE IS CONTINUOUS (DISTANCE OF VILLAGE IN KM FROM CLOSEST TOWN)[represent Cell [C]]. To explore this type of relationship we use a family of regressions called logits and probits.

. dprobit barterswap distance

Probit estimates

Number of obs = 511

LR chi2(1) = 40.18

Prob > chi2 = 0.0000

Pseudo R2 = 0.0582

Log likelihood = -324.8438

barter~ap	dF/dx	Std. Err.	z	P> z	x-bar	[95% C.I.]
distance	.0064045	.0010484	6.06	0.000	33.9198	.00435 .008459

obs. P | .5949119

pred. P | .6053342 (at x-bar)

z and P>|z| are the test of the underlying coefficient being 0

NOTE:

- Regressions are logits or (in the example above) probits
- Reading coefficient at the mean value of x. For every km farther away from the market town, the PROBABILITY of engaging in barter increases by 0.64%
- Z statistics, p value

Another way of looking at this is to compare the sample mean of distance among households that barter and that don't barter:

. sum distance if barterswap==1

Variable	Obs	Mean	Std. Dev.	Min	Max
distance	304	39.05592	25.44249	2	101

. sum distance if barterswap==0

Variable	Obs	Mean	Std. Dev.	Min	Max
distance	207	26.37681	16.89554	0	101

This simply tells you that households that bartered were farther away (39 km) from towns than households that did not barter (26 km).

[D]. TWO CATEGORICAL VARIABLES.

First, label define commands from stata to make easier the reading of 2x2 tabulations:

. label define CLOSE1b1 0 "25Km+" 1 "<25km"

. label values CLOSE CLOSE1b1

Cross tabulation to find out if people who have some education are more likely to live closer to towns:

```
. tab education CLOSE, column row chi2
```

education	CLOSE		Total
	25Km+	<25km	
no education	137	60	197
	69.54	30.46	100.00
	48.07	26.55	38.55
some education	148	166	314
	47.13	52.87	100.00
	51.93	73.45	61.45
Total	285	226	511
	55.77	44.23	100.00
	100.00	100.00	100.00

Pearson chi2(1) = 24.6447 Pr = 0.000

NOTE:

- Reading of results

[VI]. Social versus statistical significance

- Difference between the two
- Using coefficients as inputs into policy decisions/program analysis

[VII]. Limits of bivariate analysis: the analysis of Y as a function of only one X is a good starting point in the empirical analysis, but faces several problems

- What happens if the right-hand side variables interact?
 - E.g., $\text{Income} = a + \text{education} + \text{education} * \text{male}$ (i.e., education has a different effect on income depending on whether the subject is a man or a woman)
- If you need to include a square or cube term, then bivariate analysis is inadequate. More broadly, you need multivariate analysis if the relation between y and x is nonlinear (e.g., $\text{income} = a + b\text{Age} + c\text{Age}^2$, where $\text{Age}^2 = \text{Age} * \text{Age}$). U shape relations = $Y = +ax^2 + bx + c$. Inverted U-shape relation = $y = -ax^2 + bx + c$
- Omitted variable bias: causality has multiple roots (typically).
 - Even if you run an experimental research design or even if you have a truly exogenous variable, you want to control for the role of third variables
 - Think back to the Framework course: the estimated relation between income inequality (gini) (explanatory variable) and rate of economic growth rate (dependent variable) was biased because it excluded a third variable (e.g., corruption). Another example: supposed you regressed vocabulary (dependent variable) against cavities in your mouth (explanatory variable). You would probably get a big coefficient – more cavities, more vocabulary, but the relation would be driven only because of a third variable, Age, that you did not include.

Review material: effect of omitted variable bias in relation Y X1 [where X2 is omitted but where X2 is related to both Y and X1] depends on signs between:

- x1 and x2
- x2 and Y

Summary table:

If sign of X1-X2	Sign of X2-Y	Sign of indirect effect	Effect on b1 of excluding x2
+	+	+	↑
+	-	-	↓
-	-	+	↑
-	+	-	↓
* 0	+/-	0	0
* +/-	0	0	0

- For these two reasons, we move on to multivariate analysis
 - It generally provides more unbiased estimates than univariate analysis
 - Note that if X is exogenous (e.g., a randomized experiment, people winning the lottery), then the estimated parameter of X from a bivariate and a multivariate will give you the same results

[VIII]. OLS (ordinary least squares). This is the workhorse for most statistical analysis so it is important that you understand its assumption well. Make sure to look over the notes on how to read STATA output, with various examples.

- Used for bivariate and multivariate analysis
- Assumptions
 - Error term is normally distributed (Error term=residual=difference between the actual y and the predicted Y or Y hat)
 - Error terms are independent of each other
 - Error terms have zero mean and same standard deviation or spread
 - More formally,

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

Where variance of ϵ_i is homogeneous: homoskedasticity

- Violation of homoskedasticity (heteroskedasticity)
 - ➔ Will not bias the slope but
 - ➔ Will make standard errors wrong: therefore t, z, and p values will be wrong
 - ➔ Stata assumes homoskedasticity unless you tell it to correct for this
- Look at the expression above, $Y_i = \alpha + \beta X_i + \epsilon_i$. You can use the estimated coefficient, β of X_i and the constant, α , to predict, other values of x that may not be in your regression. See the two examples below for how to estimate predicted values of Y conditional on various Xs (specifically, predict weight for a person 150 cm high and 30 years of age)

METHOD #1

```
reg weight height Age
```

Source	SS	df	MS			
Model	636670.553	2	318335.277	Number of obs =	1301	
Residual	50417.9234	1298	38.8427761	F(2, 1298) =	8195.48	
				Prob > F =	0.0000	
				R-squared =	0.9266	
				Adj R-squared =	0.9265	
				Root MSE =	6.2324	
Total	687088.477	1300	528.529597			

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
height	.5566292	.0084287	66.04	0.000	.5400939	.5731646
Age	.1674249	.0173496	9.65	0.000	.1333886	.2014613
_cons	-36.68678	.841777	-43.58	0.000	-38.33817	-35.03538

```
Now substitute the constant and estimated coefficients: y = -36 + 0.55*height
[height=150] + 0.16*Age [Age=30]
. display -36+.55*150+.16*30
51.3
```

METHOD 2

```
Quietly reg weight height Age
```

```
. generate yhat=_b[_cons]+_b[height]*150+_b[Age]*30
```

```
. sum yhat
```

Variable	Obs	Mean	Std. Dev.	Min	Max
yhat	4657	51.83036	0	51.83036	51.83036.

- Regression through the origin
 - Sometimes we do not want the constant to intercept the y axis except at zero
 - In STATA we can type “reg y x, nocons”
 - Note what happens to the coefficient of the regression through the origin and one that does not go through the origin

Reasons for heteroskedasticity?

- More choices at some levels
- Outliers
- Poor model specification: for example, if the true relation is inverted U-shaped, but you use a linear relation, you will get heteroskedasticity. To convince yourself, draw a scatter plot in the form of a U where the error terms for any one x are homoskedastic. Now draw a positively-sloped line. You

will see that the error terms are now heteroskedastic, but simply because you used a poor model specification.

- Learning by doing

Detecting heteroskedasticity:

- Graphing y x. Check stata for graphing command; they are very powerful and sometimes can allow you to make the point without having to do the regression. Below are a few examples that might help you detect the relation between variables without statistical analysis:
 - ➔ scatter weight height
 - ➔ scatter weight height || lfit weight height ||, by(male, total row(1))
 - ➔ twoway histogram weight
- In stata: ➔ fit y x
 - ➔ hettest (Make sure you know how to read results. This command produces the Breusch-Pagan or Cook-Weisberg test of heteroskedasticity. The test is the ratio of two standard deviation, which is distributed as a chi2. This is a test of homoskedasticity, so if the chi2 value is large and the p>chi2 is small (say below 10%), then you reject homoskedasticity because you have heteroskedastik error terms
- Compare the standard deviation above/below a cut off (Goldfed-Quandt)

Correcting for heteroskedasticity:

- Median regression (In OLS, the line is drawn through the mean or expected values of Y (Y hat). In median regression the line is drawn through the median. Running a regression through the median corrects for the leverage power of outliers
 - ➔ In Stata: qreg y x
 - ➔ Stricter regression
- Robust standard errors:
 - ➔ weighted least squares
 - ➔ Huber-White robust weights or **robust SE**, most common correction
 - ➔ Weights used so errors become constant
 - ➔ Clustering (reg y x, cluster (z)) where z is the community level variable (e.g., village, school. Kids in same school are similar, so we cluster by school)

In OLS you have: $Y_i = B_i + B_2X_2 + e_i$

Since we have heteroskedasticity, we weigh the expression by var (ei), so

$$\frac{Y_i}{\sigma_i^2} = \frac{B_i}{\sigma_i^2} + \frac{B_2X_2}{\sigma_i^2} + \frac{e_i}{\sigma_i^2}$$

For most normal terms, it won't matter, but for outliers this procedure will bring the error term down.

- Examples from STATA
 - ➔ Quietly fit weight height
 - ➔ hettest

hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity

Ho: Constant variance

Variables: fitted values of weight

chi2(1) = 8.06

Prob > chi2 = 0.0045

The test suggests you have heteroskedasticity, therefore one option is

- reg weight height, robust

Regression with robust standard errors

Number of obs = 1301
 F(1, 1299) =13307.53
 Prob > F = 0.0000
 R-squared = 0.9214
 Root MSE = 6.4496

	weight					
	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
	.6229496	.0054001	115.36	0.000	.6123557	.6335435

Source	SS	df	MS			
Model	633053.37	1	633053.37	Number of obs =	1301	
Residual	54035.1063	1299	41.5974644	F(1, 1299) =	15218.56	
Total	687088.477	1300	528.529597	Prob > F =	0.0000	
				R-squared =	0.9214	
				Adj R-squared =	0.9213	
				Root MSE =	6.4496	

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
height	.6229496	.0050497	123.36	0.000	.6130431	.6328561
_cons	-41.87665	.6701431	-62.49	0.000	-43.19133	-40.56197

A regression corrected with the Huber-White procedure has

- Same coefficient as a regular OLS without corrections for heteroskedasticity (compare coef in two regressions above), but differences in standard errors
- Homoskedastic terms
- Can use OLS
- acknowledge heteroskedasticity, leave it in, but correct it

[IX]. Standard error of regression and implications for statistical power of estimation (Review material):

- Defined
 - T = coefficient/standard error. Note as SE increases, t or z falls, making your results less significant. To obtain statistically significant results, you want small SE
 - $SE = [(\sum(Y_i - \hat{Y})^2 / (n-k))]^{0.5}$
- Where n=sample size; k=number of explanatory variables, Y_i is actual observation, \bar{Y} is average Y, and \hat{Y} is predicted Y
- Examining the formula for the SE, make sure you understand what happens to the level of statistical significance when
 - Sample size increases/decreases
 - You have many xs in your model
 - Your X or Y lack variance (you can't tell the effect of lack of variance of X from the formula, but you should be able to tell the effect of lack of variance in Y)
 - Measurement error in Y

[X]. Reading coefficients when Y or X or both are in logs

- We use logarithms in statistical analysis
 - To correct for skewed distributions
 - To facilitate the interpretation of results. The slope of a logarithm can be read as a % Δ
- Logarithm defined: an exponent/power. Logarithmic form [$Y = \ln X$] = Exponential form [$e^Y = x$] where $x > 0$
- Sometimes we select small x if x contains many zeros, but check to see whether results are then same when you log x and log (x+ small number)
- If the Y is in log and X is in levels, we speak of log-linear model, if Y is in levels and X is in logs, we speak of the linear-log model, and if both are in logs we speak of elasticities. Make sure you know how to read coefficients when either Y or X are in logs.

Consider the following regression where the dependent variable is in logs:

$$\ln(\text{Deforestation}) = B_1 + B_2 \text{Education} + B_3 \ln(\text{Income}) + \text{error}$$

Then we can see that the coefficient B3 can be interpreted as an elasticity. If, for example, $B_3 = 0.30$, then we see that a one percent increase in income raises deforestation by 0.3 percent:

$$\frac{\Delta \ln(\text{Deforestation})}{\Delta \ln(\text{income})} = \frac{\frac{\Delta \text{Deforestation}}{\text{Deforestation}}}{\frac{\Delta \text{Income}}{\text{Income}}} = \frac{0.3}{0.01} = 0.3$$

Which implies that

$$\frac{\Delta \text{Deforestation}}{\text{Deforestation}} = 0.3 \times 0.1 = 0.003 = 0.3\%$$

Which means that deforestation rises by 0.3%

When the explanatory variable in question is not itself in logs, the interpretation is as follow:

$$\frac{\Delta \ln(\text{Deforestation})}{\text{Education}} = \frac{\Delta \ln(\text{Deforestation})}{1}$$

When education increases by one, so if the coefficient B2=0.08, then

$$\frac{\Delta \ln(\text{Deforestation})}{1} = \Delta \ln(\text{Deforestation}) = \frac{\Delta \text{Deforestation}}{\text{Deforestation}} = 0.08$$

Which means that deforestation rises by 8% when education increases by one.

Therefore, note that in the “log-log” case, the coefficient is read directly as a percentage, while in the ‘log-linear’ case we must multiply by 100 to get the percentage. In other words, in the ‘log-log’ case, the coefficient of 0.3 is interpreted as 0.3%, but in the ‘log-linear’ case, the coefficient of 0.08 is interpreted as 8% in terms of their effects on deforestation.

- Examples
 - Log linear: Reg $\ln(\text{Income}) = a + 0.2 \text{Education}$ (Interpretation: One more year of schooling increases income by 20% -- i.e., $0.2 * 100\%$)
 - Log-Log: Reg $\ln(\text{Income}) = a + \ln(\text{Education}) = a + 0.2 \ln(\text{Education})$ (Interpretation: This is an elasticity since both Y and X are in logs. A one percent increase in education increases income by 0.2%, or a doubling of education roughly increases income by 20%)
 - Linear-Log: Reg $\text{Income} = a + \ln(\text{Education}) = a + 0.20 \ln(\text{Education})$ (Interpretation: a one percent increase in education increases income by 0.20 dollars)

[XI]. Interaction variables (New material)

- So far we have assumed that variables have direct effects on y, but what happens when x variables have an interactive effect on y?
- In simple model, reg income education, but what if education has a different effect on males than on females?
- Two ways to handle the problem:
 - First method:
 - Reg income education if male==0
 - Reg income education if male==1
 - And compare coefficient
 - Example

gen income=(valuemaizeat+valuericeeat+valueyucaeat+cashSales+cashWag2wek+BarterValRec)/hhsz

. reg income education

Source	SS	df	MS	Number of obs = 690		
Model	64121.9648	1	64121.9648	F(1, 688)	=	4.98
Residual	8858810.88	688	12876.1786	Prob > F	=	0.0260
-----				R-squared	=	0.0072
Total	8922932.85	689	12950.5557	Adj R-squared	=	0.0057
-----				Root MSE	=	113.47
income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
education	4.588163	2.056027	2.23	0.026	.5513216	8.625004
_cons	46.34626	5.342496	8.68	0.000	35.85671	56.83581

. reg income education if male==1

Source	SS	df	MS	Number of obs = 517		
Model	49464.2256	1	49464.2256	F(1, 515)	=	2.96
Residual	8605706.05	515	16710.1088	Prob > F	=	0.0859
-----				R-squared	=	0.0057
Total	8655170.28	516	16773.5858	Adj R-squared	=	0.0038
-----				Root MSE	=	129.27
income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	

```

education | 4.357396 2.532626 1.72 0.086 -.6181527 9.332945
_cons | 52.01224 7.24246 7.18 0.000 37.78384 66.24064
-----+-----
. reg income education if male==0
Source | SS df MS Number of obs = 173
-----+----- F( 1, 171) = 0.93
Model | 1000.1473 1 1000.1473 Prob > F = 0.3351
Residual | 183068.872 171 1070.5782 R-squared = 0.0054
-----+----- Adj R-squared = -0.0004
Total | 184069.019 172 1070.16872 Root MSE = 32.72

-----+-----
income | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-----+-----
education | -1.769784 1.831039 -0.97 0.335 -5.384134 1.844565
_cons | 35.74402 2.890102 12.37 0.000 30.03915 41.44889
-----+-----

```

You can see that schooling has the expected positive correlation with income, but only with men; with women, schooling does not correlate significantly with income. One problem with this approach is the loss of the excluded category in the estimation (e.g., when estimating the effect of schooling on income for men, you exclude women).

→ Second method

- Generate an interaction variable between education and male
- gen edumale=education*male
- reg income education edumale male

```

. reg income education edumale male
Source | SS df MS Number of obs = 690
-----+----- F( 3, 686) = 3.49
Model | 134157.921 3 44719.307 Prob > F = 0.0155
Residual | 8788774.93 686 12811.6253 R-squared = 0.0150
-----+----- Adj R-squared = 0.0107
Total | 8922932.85 689 12950.5557 Root MSE = 113.19

-----+-----
income | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-----+-----
education | -1.769784 6.334181 -0.28 0.780 -14.20649 10.66692
edumale | 6.12718 6.711155 0.91 0.362 -7.04969 19.30405
male | 16.26822 11.83945 1.37 0.170 -6.977689 39.51413
_cons | 35.74402 9.997837 3.58 0.000 16.11398 55.37405
-----+-----

```

Using the second equation answer the following two questions: (a) what is the impact of an additional year of schooling for women? (b) for men? How do these results compare with the results from the first method?

[XII]. Joint significance: Sometimes when running a multivariate regression we want to estimate the significance of a group of variables. After the regression type “test x1 x2” i.e., all the variables for which you want to estimate the test of joint significance. The test compares the R2 of the restricted and unrestricted model.

```
. reg income writing education math speakSpanish hhsizes walkaccess Age male
```

```

Source | SS df MS Number of obs = 690
-----+----- F( 8, 681) = 6.25
Model | 610509.036 8 76313.6295 Prob > F = 0.0000
Residual | 8312423.81 681 12206.2024 R-squared = 0.0684
-----+----- Adj R-squared = 0.0575
Total | 8922932.85 689 12950.5557 Root MSE = 110.48

-----+-----
income | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-----+-----
writing | -18.11829 9.797733 -1.85 0.065 -37.35568 1.119106
education | 7.479641 3.428298 2.18 0.029 .7483366 14.21094
math | -3.358189 5.114605 -0.66 0.512 -13.40048 6.684101
speakSpanish | 2.969216 7.779081 0.38 0.703 -12.30465 18.24308
hhsizes | -9.293917 1.685388 -5.51 0.000 -12.6031 -5.984737

```

walkaccess		-.6318018	.4463466	-1.42	0.157	-1.508183	.2445789
Age		-.2749659	.3798628	-0.72	0.469	-1.020809	.4708771
male		20.58811	10.79628	1.91	0.057	-.6098902	41.78611
_cons		109.5877	19.3278	5.67	0.000	71.63846	147.5369

```
. test writing education math speakSpanish
```

```
( 1)  writing = 0
( 2)  education = 0
( 3)  math = 0
( 4)  speakSpanish = 0
```

```
F( 4, 681) = 1.55
Prob > F = 0.1852
```

The tests of joint significance can be applied to test for non-linearities. For example,

```
reg y Age Age2 hhsz education
```

is a perfectly reasonable model, except we don't know whether the relation between y and Age and Age2 is non-linear. To test for non-linearity, we would do as follow

```
reg y Age Age2 hhsz education
test Age Age2
```

The second line would tell us whether Age and Age2, jointly, are strong covariates of y .

Table 1. Definition and summary statistic of variables used in statistical analysis

Name	Definition	N	Mean	Std Dev
Dependent variables for adults:				
<i>Short-run nutritional status:</i>				
BMI	Body-mass index; kg/m ²	556	23.173	2.463
ZAM	Sex and age-standardized z score of mid-arm muscle area following Frisancho's norm {Frisancho 1990 10 /id}	559	-.831	.788
ZSF	Age and sex-standardized z scores of sum of triceps and subscapular skinfold thickness following Frisancho's norm {Frisancho 1990 10 /id}	561	-.717	.492
ZWT	Age and sex-standardized z score of weight for age following norms of the National Center for Health Statistics {Hamill, Drizd, et al. 1979 5227 /id}	558	-1.047	.540
<i>Perceived illness:</i>				
Days ill	# days ill last week (up to 3 ailments); 0-21	561	2.071	3.234
Days bed	# days in bed last week; 0-7	562	.700	1.549
<i>Earnings:</i>				
Earnings	(Cash earnings from wage labor + sale of goods)/household size measured with adult equivalents; bolivianos/person. 1 US dollar = 6.31 bolivianos in 2001-2002; in regression entered as logarithm	559	42.920	105.193
Dependent variables for children (1-12):				
<i>Short-run nutritional status:</i>				
ZAM	Sex and age-standardized z score of mid-arm muscle area following Frisancho's norm {Frisancho 1990 10 /id}	514	-.695	.857
ZSF	Age and sex-standardized z scores of sum of triceps and subscapular skinfold thickness following Frisancho's norm {Frisancho 1990 10 /id}	514	-.610	.618
ZWT	Age and sex-standardized z score of weight for age following norms of the National Center for Health Statistics {Hamill, Drizd, et al. 1979 5227 /id}	514	-1.112	1.102
<i>Perceived illness:</i>				
Days ill	# days ill last week (up to 3 ailments); 0-21	514	3.178	3.771
Days bed	# days in bed last week; 0-7	514	1.305	1.976

Table 1. Definition and summary statistic of variables used in statistical analysis

Name	Definition	N	Mean	Std Dev
Explanatory variables for adults:				
<i>Traditional culture:</i>				
Culture	Total score on Likert scale to measure valuation of own culture; 9 questions; 1=agree; 1=disagree; 3=indifferent or does not know.	558	19.734	3.136
Generosity	Total episodes of generosity in last week toward other Tsimane' per adult equivalents. See text.	559	4.110	4.751
Monolingual	Monolingual in Tsimane' (1=yes; 0=bilingual in Tsimane' and Spanish).	554	.341	.474
Wealth Traditional	Village-gate value of 8 traditional physical assets per adult equivalent in <u>bolivianos</u> ; in Table 4 entered in logarithms	559	313.608	433.276
Barter	Value in <u>bolivianos</u> of goods received in barter in last 2 weeks per adult equivalent	559	13.073	26.216
Traditional	Principal component score of above variables under section "traditional culture"	548	-1.09e-09	1.391
<i>Modern culture:</i>				
Wealth modern	Village-gate value of 11 modern physical assets per adult equivalent in <u>bolivianos</u> ; in Table 4 in logarithms	559	319.345	257.461
Education	Maximum school attainment of subject	554	1.517	2.198
Writing	Writing skills in Spanish based on test; 0=cannot; 1=with difficulty; 2=well	554	.543	.799
Math	Score in math test; 4 questions to test four basic arithmetic operations. Score: 0-4	561	.841	1.323
Reading	Reading skills in Spanish based on test; 0=cannot; 1=with difficulty; 2=well.	554	.519	.811
Modern	Principal component score of above variables under section "modern culture"	548	1.90e-09	1.064
<i>Controls:</i>				
Age	Age of subject in years	562	34.297	12.608
Male	Sex of subject; 1=male; 0=female	562	.693	.461
ZHT	Sex and age-standardized z score of height for age using Frisancho's norms {Frisancho 1990 10 /id}	559	-1.798	.710
Encroacher	Instrument to correct for selectivity in earnings; 1=household in contact with encroachers in last month, 0 otherwise	562	.733	.442
Size	Household size with adult equivalents	562	4.063	1.709
Distance	Village-to-town walking time in hours	38	13.065	12.484