



Getting Started in Linear Regression

(ver. 3.5 *beta*)

Oscar Torres-Reyna

Data Consultant

otorres@princeton.edu



List of topics

☐ Regression:

- ✓ Overview and basic setting
- ✓ Correlation matrix
- ✓ Output interpretation (what to look for)
- ✓ Graph matrix
- ✓ Saving regression coefficients
- ✓ F-test
- ✓ Testing for linearity
- ✓ Testing for normality

☐ Regression:

- ✓ Testing for homoskedasticity
- ✓ Testing for omitted-variable bias
- ✓ Testing for multicollinearity
- ✓ Robust standard errors
- ✓ Specification error
- ✓ Outliers
- ✓ Interaction terms
- ✓ Publishing regression table

☐ Useful sites (links only)

- ✓ *Is my model OK?*
- ✓ *I can't read the output of my model!!!*
- ✓ Topics in Statistics
- ✓ Recommended books

In this section we will explore some basics of regression analysis.

We will run a multivariate regression and some diagnostics :

- General setting and output interpretation (what to look for)
- Normality
- Linearity/functional form
- Homoskedasticity/heteroskedasticity
- Robust standard errors
- Omitted variable bias/specification error
- Outliers
- F -test
- Interaction terms

The main references/sources for this section are:

- Stock, James and Mark Watson, *Introduction to Econometrics*, 2003
- Hamilton, Lawrence, *Statistics with Stata (updated for version 9)*, 2006
- The UCLA online tutorial <http://www.ats.ucla.edu/stat/stata/>

We use regression to estimate the unknown effect of changing one variable over another (Stock and Watson, 2003, ch. 4)

When we run a regression we assume a linear relationship between two variables (i.e. X and Y). Technically, it estimates how much Y changes when X changes one unit.

In Stata we use the command `regress`, type:

```
regress [dependent variable] [independent variable(s)]
```

```
regress y x
```

In a multivariate setting we type:

```
regress y x1 x2 x3 ...
```

Before running a regression it is recommended to have a clear idea of what you are trying to estimate (i.e. which are your dependent and independent variables).

A regression makes sense only if there is a sound theory behind it.

Regression: a practical approach (overview) cont.

For this section I am going to use data and examples from the book *Statistics with Stata (updated for version 9)* by Lawrence C. Hamilton (chapter 6). [Click here](#) to download the data or search for it at <http://www.duxbury.com/highered/>.

Using the file `states.dta` (educational data for the U.S.).

```
. use states
(U.S. states data 1990-91)

. describe

Contains data from states.dta
  obs:      51          U.S. states data 1990-91
  vars:     21          14 Sep 2003 18:34
  size:    4,386 (99.9% of memory free)
```

variable name	storage type	display format	value label	variable label
state	str20	%20s		State
region	byte	%9.0g	region	Geographical region
pop	float	%9.0g		1990 population
area	float	%9.0g		Land area, square miles
density	float	%7.2f		People per square mile
metro	float	%5.1f		Metropolitan area population, %
waste	float	%5.2f		Per capita solid waste, tons
energy	int	%8.0g		Per capita energy consumed, Btu
miles	float	%8.0g		Per capita miles/year, 1,000
toxic	float	%5.2f		Per capita toxics released, lbs
green	float	%5.2f		Per capita greenhouse gas, tons
house	byte	%8.0g		House '91 environ. voting, %
senate	byte	%8.0g		Senate '91 environ. voting, %
csat	int	%9.0g		Mean composite SAT score
vsat	int	%8.0g		Mean verbal SAT score
msat	int	%8.0g		Mean math SAT score
percent	byte	%9.0g		% HS graduates taking SAT
expense	int	%9.0g		Per pupil expenditures prim&sec
income	double	%10.0g		Median household income, \$1,000
high	float	%9.0g		% adults HS diploma
college	float	%9.0g		% adults college degree

```
sorted by: state
```

Regression: a practical approach (setting)

Starting question: *Are SAT scores higher in states that spend more money on education controlling by other factors?*

- Dependent (or predicted, Y) variable – SAT scores, variable `csat` in dataset
- Independent (or predictor, X) variable(s) – Expenditures on education, variable `expense` in dataset. Other variables `percent`, `income`, `high`, `college`.

Here is a general description of the variables in the model

```
. describe csat expense percent income high college
```

variable name	storage type	display format	value label	variable label
csat	int	%9.0g		Mean composite SAT score
expense	int	%9.0g		Per pupil expenditures prim&sec
percent	byte	%9.0g		% HS graduates taking SAT
income	double	%10.0g		Median household income, \$1,000
high	float	%9.0g		% adults HS diploma
college	float	%9.0g		% adults college degree

```
. summarize csat expense percent income high college
```

Variable	Obs	Mean	Std. Dev.	Min	Max
csat	51	944.098	66.93497	832	1093
expense	51	5235.961	1401.155	2960	9259
percent	51	35.76471	26.19281	4	81
income	51	33.95657	6.423134	23.465	48.618
high	51	76.26078	5.588741	64.3	86.6
college	51	20.02157	4.16578	12.3	33.3

```
. corr csat expense percent income high college
(obs=51)
```

	csat	expense	percent	income	high	college
csat	1.0000					
expense	-0.4663	1.0000				
percent	-0.8758	0.6509	1.0000			
income	-0.4713	0.6784	0.6733	1.0000		
high	0.0858	0.3133	0.1413	0.5099	1.0000	
college	-0.3729	0.6400	0.6091	0.7234	0.5319	1.0000

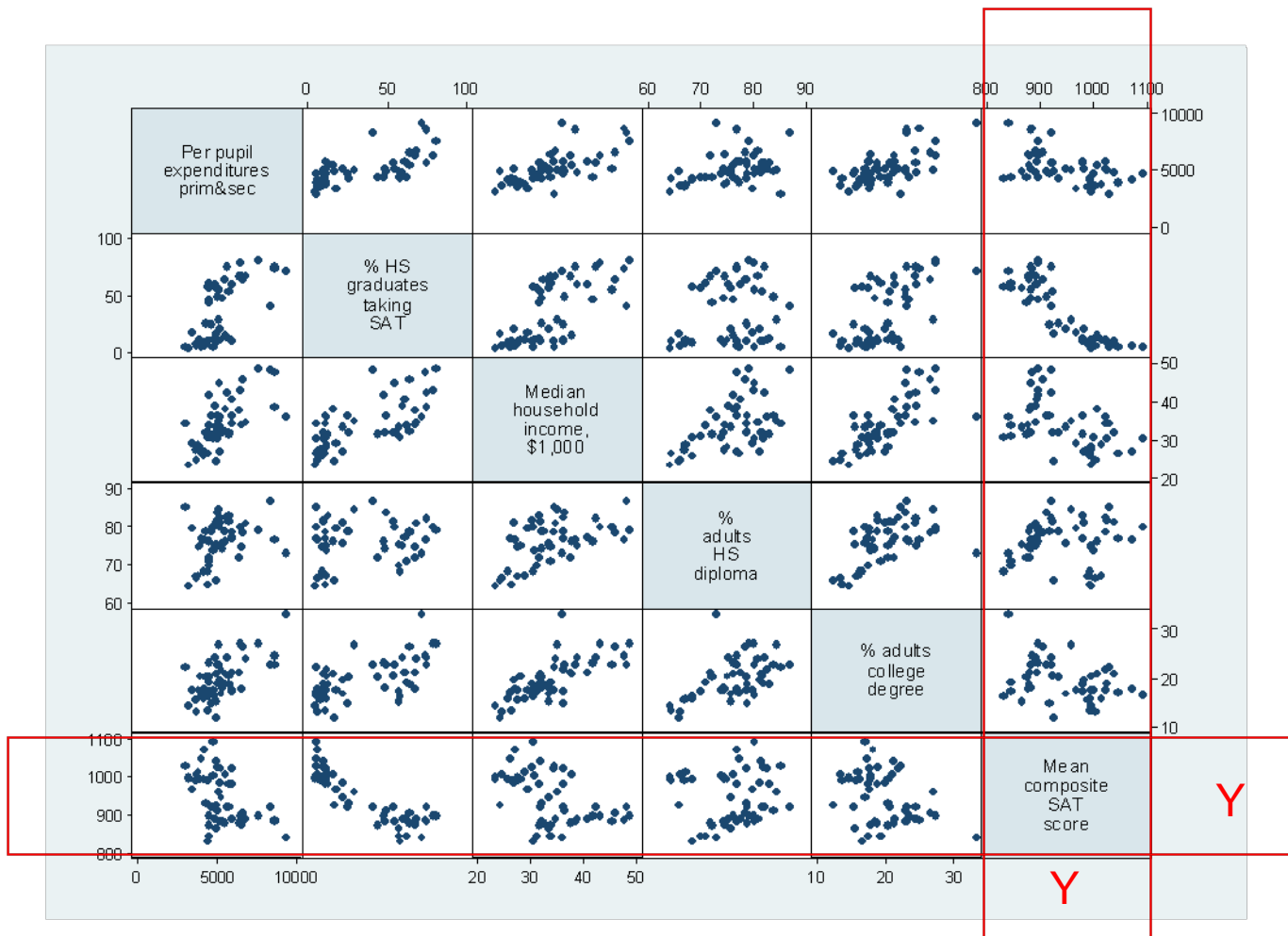
This is a correlation matrix for all variables in the model. Numbers are Pearson correlation coefficients, go from -1 to 1. Closer to 1 means strong correlation. A negative value indicates an inverse relationship (roughly, when one goes up the other goes down).



Regression: graph matrix

Before running a regression is always recommended to graph dependent and independent variables to explore their relationship. Command `graph matrix` produces a series of scatterplots for all variables. Type:

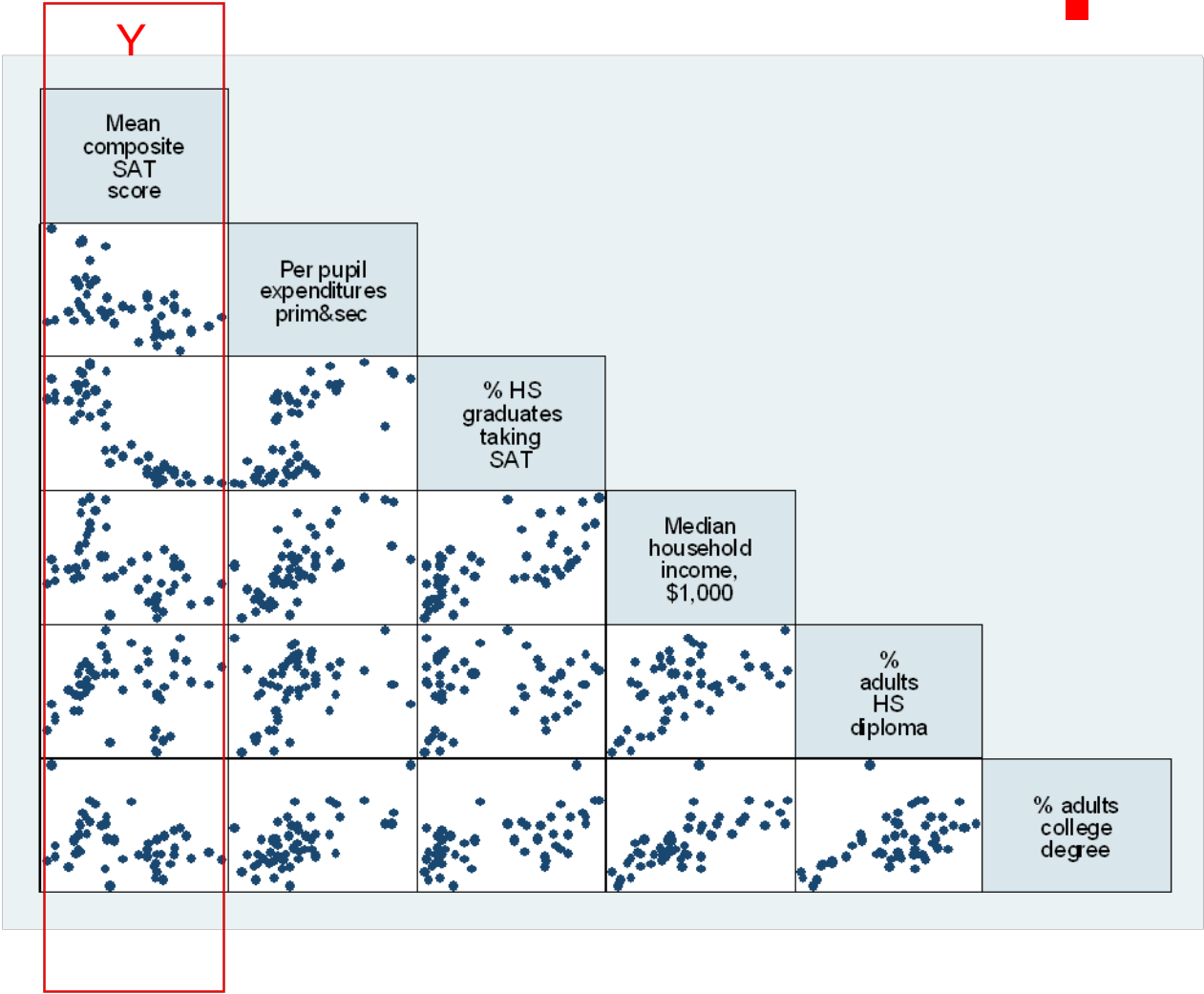
```
graph matrix expense percent income high college csat
```



Regression: graph matrix

Here is another option for the graph.

```
graph matrix csat expense percent income high college, half  
maxis(ylabel(none) xlabel(none))
```



Regression: what to look for

Lets run the regression:

```
regress csat expense percent income high college, robust
```

Dependent variable (Y)

Independent variables (X)

```
. regress csat expense percent income high college, robust
Linear regression                               Number of obs =      51
                                                F( 5, 45) =      50.90
                                                Prob > F =      0.0000
                                                R-squared =      0.8243
                                                Root MSE =      29.571
```

csat	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
expense	.0033528	.004781	0.70	0.487	-.0062766	.0129823
percent	-2.618177	.2288594	-11.44	0.000	-3.079123	-2.15723
income	.1055853	1.207246	0.09	0.931	-2.325933	2.537104
high	1.630841	.943318	1.73	0.091	-.2690989	3.530781
college	2.030894	2.113792	0.96	0.342	-2.226502	6.28829
_cons	851.5649	57.28743	14.86	0.000	736.1821	966.9477

1

This is the p-value of the model. It indicates the reliability of X to predict Y. Usually we need a p-value lower than 0.05 to show a statistically significant relationship between X and Y.

2

R-square shows the amount of variance of Y explained by X. In this case the model explains 82.43% of the variance in SAT scores.

3

Adj R-square shows the same as R-sqr but adjusted by the number of cases and number of variables. When the number of variables is small and the number of cases is very large then Adj R-square is closer to R-square. This provides a more honest association between X and Y.

$csat = 851.56 + 0.003*expense$
 $- 2.62*percent + 0.11*income + 1.63*high$
 $+ 2.03*college$

The t-values test the hypothesis that the coefficient is different from 0. To reject this, you need a t-value greater than 1.96 (at 0.05 confidence). You can get the t-values by dividing the coefficient by its standard error. The t-values also show the importance of a variable in the model. In this case, percent is the most important.

5

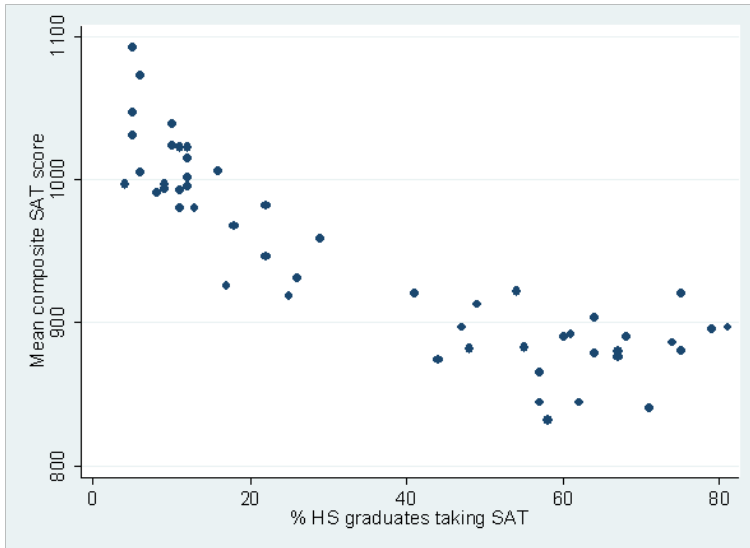
These are two-tail p-values for each coefficient. It tests the hypothesis that the coefficient is different from 0. To reject this, the p-value has to be lower than 0.05 (you could choose also an alpha of 0.01). In this case, expense, income, and college are not statistically significant in explaining SAT; high is almost significant at 0.10. Percent is the only variable that has some significant impact on SAT (its coefficient is different from 0)

4

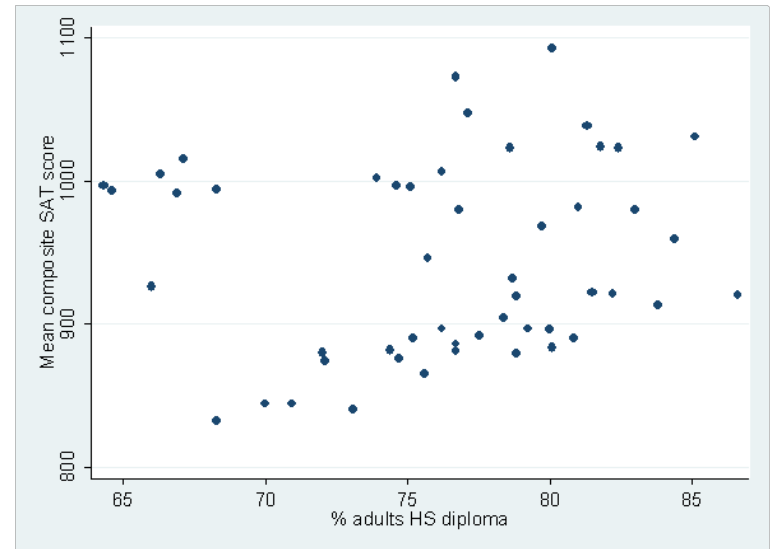
Regression: exploring relationships

Given the previous results, we need to do some adjustments since only one variable was significant. Lets explore further the relationship between `csat` and `percent`, and `high`.

```
scatter csat percent
```



```
scatter csat high
```



There seem to be a curvilinear relationship between `csat` and `percent`, and slightly linear between `csat` and `high`. Whenever we find polynomial relationships (curves) we need to add a square (or some other higher power) version of the variable, in this case `percent square` will suffice.

```
generate percent2 = percent^2
```

Now the model will look like this

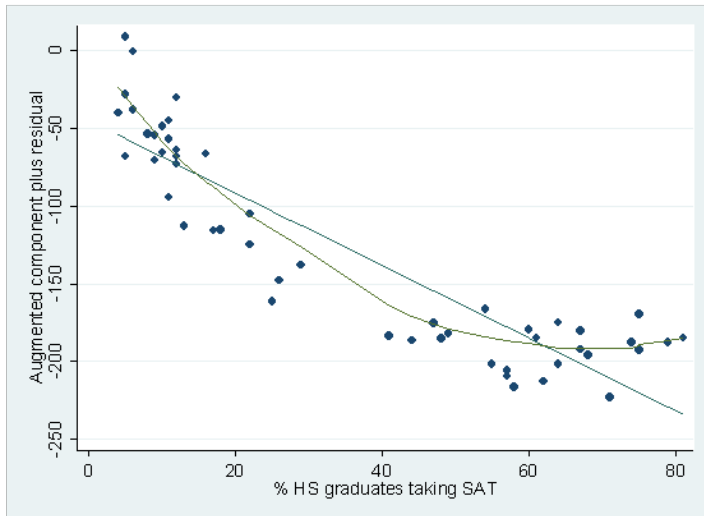
```
regress csat percent percent2 high
```

Regression: functional form/linearity

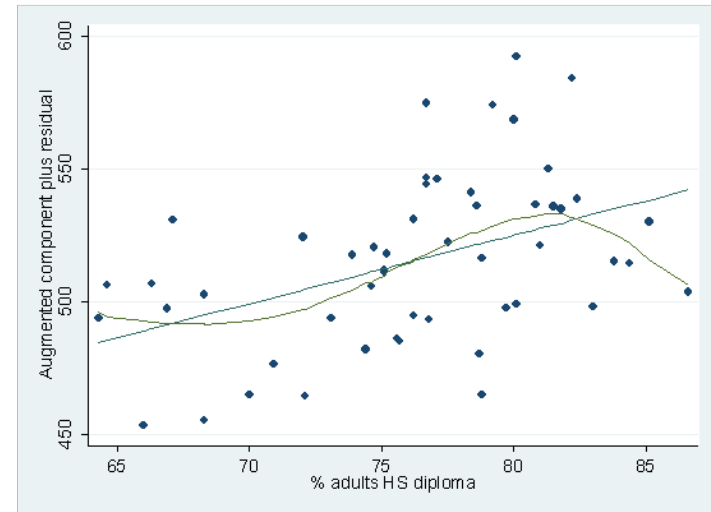
As a footnote, another graphical way to explore a possible linear relationship between variables or to detect nonlinearity to define a functional form is by using the command `acprplot` (augmented component-plus-residual plot). Right after running a the regression:

```
regress csat percent high /* Notice we do not include percent2 */
```

```
acprplot percent, lowess
```



```
acprplot high, lowess
```



The option `lowess` (locally weighted scatterplot smoothing) draw the observed pattern in the data to help identify nonlinearities. `Percent` shows a quadratic relation, it makes sense to add a square version of it. `High` shows a polynomial pattern as well but goes around the regression line (except on the right). We could keep it linear for now.

Form more details see <http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter2/statareg2.htm>, and/or type `help acprplot` and `help lowess`.

Regression: F-test

Before we continue let's take another look at the original regression and run some individual tests on its coefficients.

We have two types of tests with the regression model: *F*-test, which tests the overall fit of the model (all coefficients different from 0) and *t*-test (individual coefficients different from 0). You can customize your tests to check for other possible situations, like two coefficients jointly different from 0. In the regression, two variables related to educational attainment were not significant. We could, however, test whether these two have no effect on SAT scores (see Hamilton, 2006, p.175). Let's run the original regression again:

```
quietly regress csat expense percent income high college
```

Note 'quietly' suppress the regression output

To test the null hypothesis that *both* coefficients do not have any effect on *csat*, type:

```
test high college
```

```
. test high college
( 1)  high = 0
( 2)  college = 0

      F( 2, 45) =    3.32
      Prob > F =    0.0451
```

The p-value is 0.0451, under the 0.05 usual threshold (95% confidence) so we conclude that *both* variables have indeed some effect on SAT. In a way, this is saying that both have similar effect or measuring the same thing (which could suggest multicollinearity). We could keep *high* since it was borderline significant.

Some other possible tests are (see Hamilton, 2006, p.176):

```
test income = 1
```

```
test high = college
```

```
test income = (high + college)/100
```

Note: Not to be confused with `ttest`. Type `help test` and `help ttest` for more details

Regression: output

Lets try the new model. It has now a higher R-squared (0.92) and all the variables are significant.

```
. regress csat percent percent2 high
```

Source	SS	df	MS	Number of obs =	51
Model	207225.103	3	69075.0343	F(3, 47) =	193.37
Residual	16789.4069	47	357.221424	Prob > F =	0.0000
				R-squared =	0.9251
				Adj R-squared =	0.9203
Total	224014.51	50	4480.2902	Root MSE =	18.9

csat	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
percent	-6.520312	.5095805	-12.80	0.000	-7.545455 -5.495168
percent2	.0536555	.0063678	8.43	0.000	.0408452 .0664659
high	2.986509	.4857502	6.15	0.000	2.009305 3.963712
_cons	844.8207	36.63387	23.06	0.000	771.1228 918.5185

The new equation is:

$$csat = 844.82 - 6.52 * percent + 0.05 * percent2 + 2.98 * high$$

Percent's coefficient is -6.52. So, if percent increases by one unit, csat will decrease by 6.52 units. With a statistically significant p-value of 0.000 (which means that -6.52 is statistically different from 0), percent has an important impact on csat controlling by other variables (holding them constant). You could read percent2 (which explains the upward effect) the same way. The net effect of percent is the difference between both coefficients (which is still negative).

High's coefficient is 2.98. So, if high increases by one unit, csat will increase by 2.98 units.

The constant 844.82 means that if all variables are 0, the average csat score would be 844.82. It is where the regression line crosses the Y axis.

Regression: saving regression coefficients/getting predicted values

Stata temporarily stores the coefficients as `_b[varname]`, so if you type:

You can save the coefficients as variables by typing:

```
gen percent_coeff = _b[percent]
gen percent_coeff = _b[percent2]
gen high_coeff = _b[high]
gen constant_coeff = _b[_cons]
```



```
. display _b[percent]
-6.5203116
. display _b[percent2]
.05365555
. display _b[high]
2.9865088
. display _b[_cons]
844.82067
```

How good the model is will depend on how well it predicts Y and on the validity of the tests.

There are two ways to generate the *predicted values of Y* (usually called *\hat{Y}*) given the model:

Option A, using `generate` after running the regression:

```
generate csat_predict = _b[_cons] + _b[percent]*percent + _b[percent2]*percent2 + _b[high]*high
```

Option B, using `predict` immediately after running the regression:

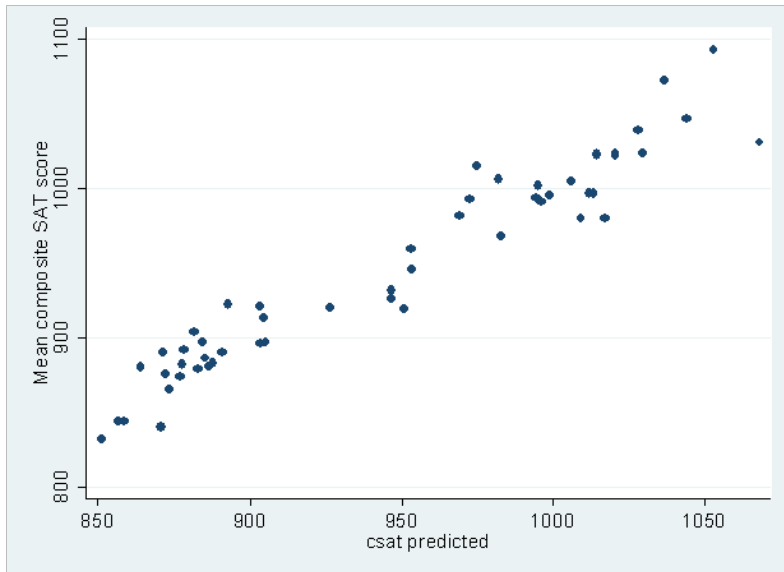
```
predict csat_predict
label variable csat_predict "csat predicted"
```

```
. predict csat_predict
(option xb assumed; fitted values)
. label variable csat_predict "csat predicted"
```

Regression: observed vs. predicted values

Now lets see how well we did, type

```
scatter csat csat_predict
```



We should expect a 45 degree pattern in the data. Y-axis is the observed data and x-axis the predicted data (\hat{Y}). In this case the model seems to be doing a good job in predicting `csat`

Regression: testing for normality

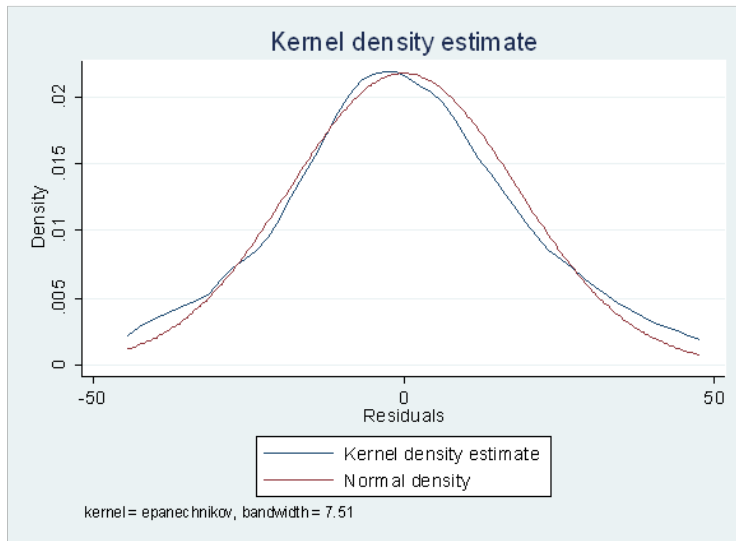
A main assumption of the regression model (OLS) that guarantee the validity of all tests (p, t and F) is that residuals behave 'normal'. Residuals (here indicated by the letter "e") are the difference between the observed values (Y) and the predicted values (Yhat): $e = Y - \hat{Y}$.

In Stata you type: `predict e, resid`

It will generate a variable called "e" (residuals).

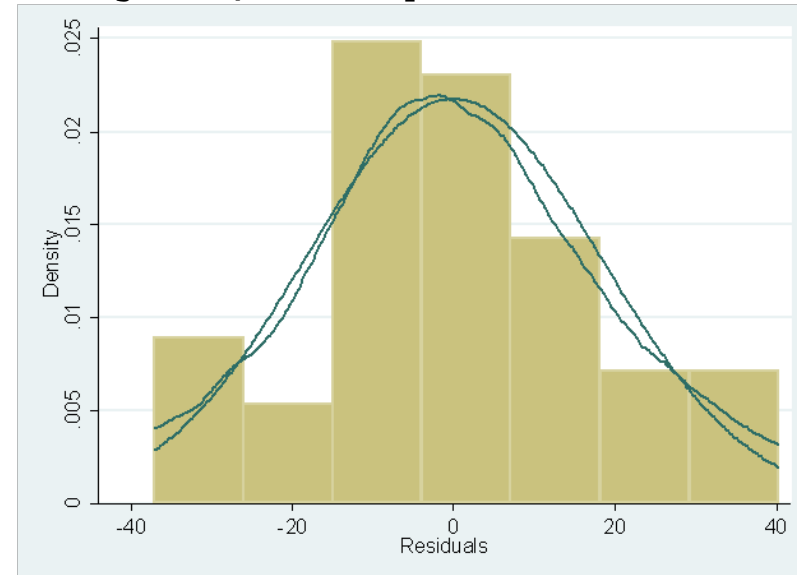
Three graphs will help us check for normality in the residuals: `kdensity`, `pnorm` and `qnorm`.

`kdensity e, normal`



A kernel density plot produces a kind of histogram for the residuals, the option `normal` overlays a normal distribution to compare. Here residuals seem to follow a normal distribution. Below is an example using `histogram`.

`histogram e, kdensity normal`

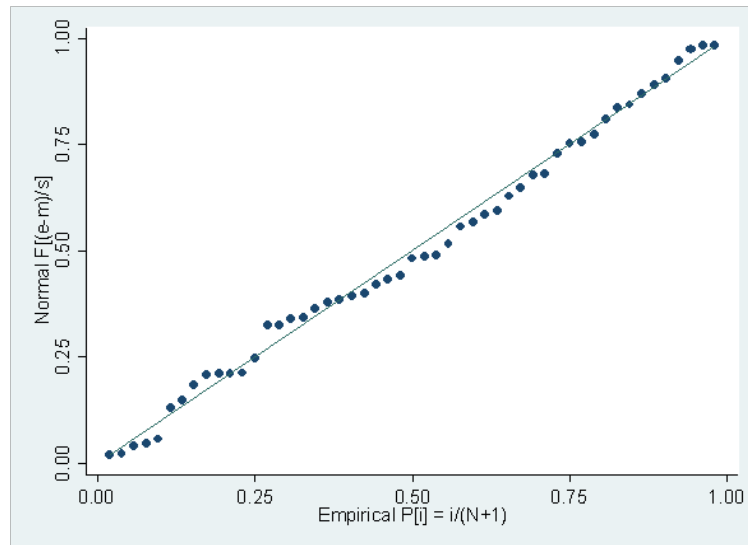


If residuals do not follow a 'normal' pattern then you should check for omitted variables, model specification, linearity, functional forms. In sum, you may need to reassess your model/theory. In practice normality does not represent much of a problem when dealing with really big samples.

Regression: testing for normality

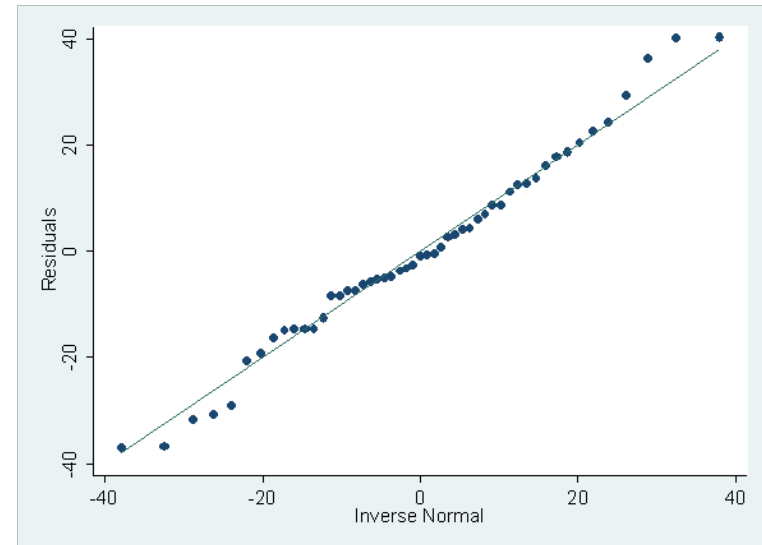
Standardize normal probability plot (`pnorm`) checks for non-normality in the middle range of residuals. Again, slightly off the line but looks ok.

`pnorm e`



Quintile-normal plots (`qnorm`) check for non-normality in the extremes of the data (tails). It plots quintiles of residuals vs quintiles of a normal distribution. Tails are a bit off the normal.

`qnorm e`



A non-graphical test is the Shapiro-Wilk test for normality. It tests the hypothesis that the distribution is normal, in this case the null hypothesis is that the distribution of the residuals is normal. Type

`swilk e`

```
. swilk e
```

variable	obs	w	V	Z	Prob>z
e	51	0.98238	0.842	-0.368	0.64349

The null hypothesis is that the distribution of the residuals is normal, here the p-value is 0.64 (way over the usual 0.05 threshold) therefore we failed to reject the null. We conclude then that residuals are normally distributed.

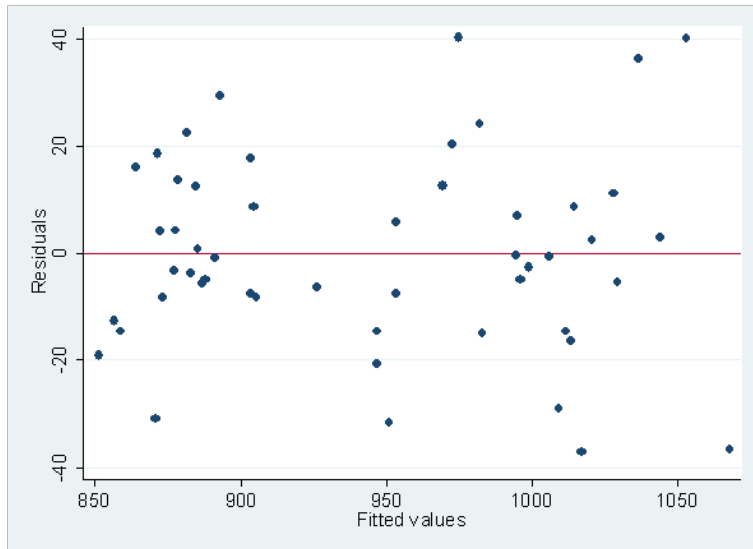
Regression: testing for homoskedasticity

Another important assumption is that the variance in the residuals has to be homoskedastic, which means constant. Residuals cannot varied for lower of higher values of X (i.e. fitted values of Y since $Y=Xb$). A definition:

“The error term $[e]$ is homoskedastic if the variance of the conditional distribution of $[e_i]$ given X_i [$\text{var}(e_i|X_i)$], is constant for $i=1 \dots n$, and in particular does not depend on x ; otherwise, the error term is heteroskedastic” (Stock and Watson, 2003, p.126)

When plotting residuals vs. predicted values (\hat{Y}) we *should not observe* any pattern at all. In Stata we do this using `rvfplot` right after running the regression, it will automatically draw a scatterplot between residuals and predicted values; and `hettest` to produce a non-graphical test.

`rvfplot, yline(0)`



`estat hettest`

```
. estat hettest  
  
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity  
Ho: Constant variance  
Variables: fitted values of csat  
  
chi2(1) = 4.86  
Prob > chi2 = 0.0274
```



Residuals seem to slightly expand at higher levels of \hat{Y} .



This is the Breusch-Pagan test for heteroskedasticity. The null hypothesis is that residuals are homoskedastic. Here we reject the null and concluded that residuals are heteroskedastic.

These two tests suggest the presence of heteroskedasticity in our model. The problem with this is that we may have the wrong estimates of the standard errors for the coefficients and therefore their t-values.

By default Stata assumes homoskedastic standard errors, so we need to adjust our model to account for heteroskedasticity. To do this we use the option `robust` in the `regress` command.

```
regress csat percent percent2 high, robust
```

See the next slide for results

Regression: robust standard errors

To run a regression with robust standard errors type:

```
regress csat percent percent2 high, robust
```

Notice the difference in the standard errors and the t-values. Following Stock and Watson, as a rule-of-thumb, you should always assume heteroskedasticity in your model and use robust standard errors by adding the option `robust` (or `r` for short) to the regression command (see Stock and Watson, 2003, chapter 4)

```
. regress csat percent percent2 high, robust
```

Linear regression

Number of obs =	51
F(3, 47) =	160.90
Prob > F =	0.0000
R-squared =	0.9251
Root MSE =	18.9

csat	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
percent	-6.520312	.4934097	-13.21	0.000	-7.512924	-5.527699
percent2	.0536555	.0056491	9.50	0.000	.042291	.0650201
high	2.986509	.54564	5.47	0.000	1.888823	4.084195
_cons	844.8207	38.8214	21.76	0.000	766.7221	922.9192

Regression: omitted-variable test

How do we know we have included all variables we need to explain Y?

Testing for omitted variable bias is important for our model since it is related to the assumption that the error term and the independent variables in the model are not correlated ($E(e|X) = 0$)

If we are missing one variable in our model and “[1] is correlated with the included regressor; and [(2)] the omitted variable is a determinant of the dependent variable” (Stock and Watson, 2003, p.144), then our regression coefficients are inconsistent.

In Stata we test for omitted-variable bias using the `ovtest` command. After running the regression type:

```
ovtest
```

```
. ovtest
Ramsey RESET test using powers of the fitted values of csat
Ho: model has no omitted variables
      F(3, 44) =      1.48
      Prob > F =      0.2319
```

The null hypothesis is that the model does not have omitted-variables bias, the p-value is 0.2319 higher than the usual threshold of 0.05, so we fail to reject the null and conclude that we do not need more variables.

Regression: specification error

Another command to test model specification is `linktest`. It basically checks whether we need more variables in our model by running a new regression with the observed Y (`csat`) against `Yhat` (`csat_predicted`) and `Yhat-squared` as independent variables¹.

The thing to look for here is the significance of `_hatsq`. The null hypothesis is that there is no specification error. If the p-value of `_hatsq` is not significant then we fail to reject the null and conclude that our model is correctly specified.

```
. linktest
```

Source	SS	df	MS			
Model	207270.449	2	103635.225	Number of obs =	51	
Residual	16744.0604	48	348.834592	F(2, 48) =	297.09	
				Prob > F =	0.0000	
				R-squared =	0.9253	
				Adj R-squared =	0.9221	
				Root MSE =	18.677	
Total	224014.51	50	4480.2902			

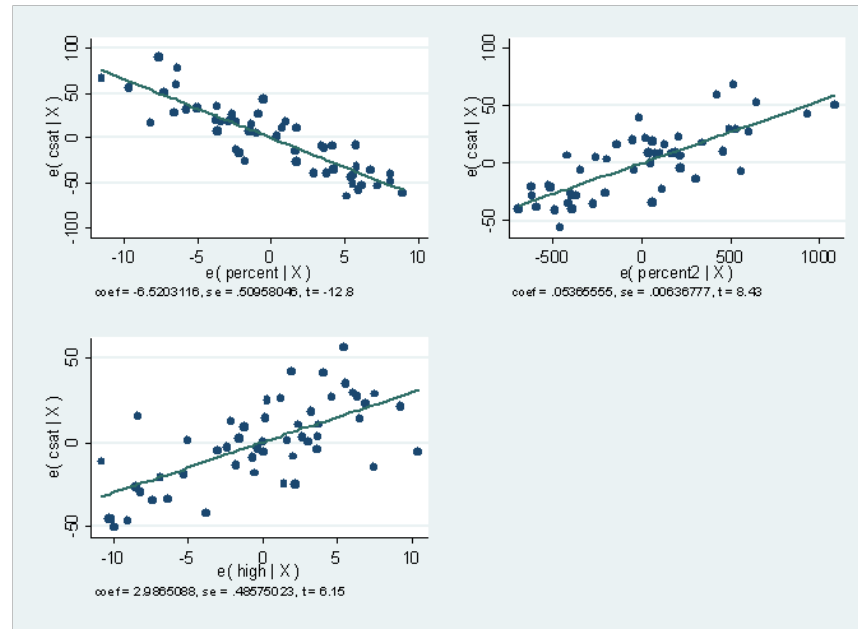
csat	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_hat	1.588861	1.633699	0.97	0.336	-1.69591	4.873632
_hatsq	-.00031	.0008597	-0.36	0.720	-.0020384	.0014185
_cons	-278.4089	773.132	-0.36	0.720	-1832.895	1276.077

¹ For more details see <http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter2/statareg2.htm>, and/or type `help linktest`.

Regression: outliers

To check for outliers we use the `avplots` command (added-variable plots). Outliers are data points with extreme values that could have a negative effect on our estimators. After running the regression type:

```
avplots
```



These plots regress each variable against all others, notice the coefficients on each. All data points seem to be in range, no outliers observed.

For more details and tests on this and influential and leverage variables please check <http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter2/statareg2.htm>

Also type `help diagplots` in the Stata command window.

Regression: multicollinearity

An important assumption for the multiple regression model is that independent variables are *not perfectly multicollinear*. This is, one regressor should not be a linear function of another. When multicollinearity is present, Stata will drop one of the variables to avoid a division by zero in the OLS procedure (see Stock and Watson, 2003, chapter 5). A mayor problem with multicollinearity is that *standand errors may be inflated*. The Stata command to check for multicollinearity is `vif` (variance inflation factor). Right after running the regression type:

```
. vif
```

variable	VIF	1/VIF
percent	24.94	0.040103
percent2	24.78	0.040354
high	1.03	0.969423
Mean VIF	16.92	

A $vif > 10$ or a $1/vif < 0.10$ indicates trouble. We know that `percent` and `percent2` are related since one is the square of the other. They are ok since `percent` has a quadratic relationship with `Y`. `High` has a `vif` of 1.03 and `1/vif` of 0.96 so we are ok here.

Lets run another regression and get the `vif`.

```
quietly regress csat expense percent income high college
```

```
. vif
```

variable	VIF	1/VIF
income	3.21	0.311756
college	2.73	0.365683
percent	2.53	0.395603
expense	2.24	0.445673
high	1.76	0.568732
Mean VIF	2.49	

We do not observe multicollinearity problems here. All `vifs` are under 10 .

Regression: publishing regression output (outreg2)

Once you define your final model, you can export your regression results using either your log file or the option outreg2. For the log you just open it using any word processor and copy-and-paste the regression table into excel or word. The command outreg2 gives you the type of presentation you see in scholar's papers. Let's say the final regression is

```
regress csat percent percent2 high
```

After running the regression type the following if you want to export the results to excel*

```
outreg2 using results, bdec(2) tdec(2) rdec(2) addec(2) alpha(0.001, 0.01, 0.05) addstat(Adj. R-squared, e(r2_a)) excel
```

Or this if you want to export to word

```
outreg2 using results, bdec(2) tdec(2) rdec(2) addec(2) alpha(0.001, 0.01, 0.05) addstat(Adj. R-squared, e(r2_a)) word
```

You will see this in Stata's output window

For excel

```
. outreg2 using results, bdec(2) tdec(2) rdec(2) addec(2) alpha(0.001, 0.01, 0.05) addstat(Adj. R-squared, e(r2_a)) excel
> (r2_a)) excel
"results.xml"
seeout
```

Click here to see the output, a excel/word window will open

For word

```
. outreg2 using results, bdec(2) tdec(2) rdec(2) addec(2) alpha(0.001, 0.01, 0.05) addstat(Adj. R-squared, e(r2_a)) word
> (r2_a)) word
"results.rtf"
seeout
```

Click on seeout to browse the results

Name of the file for the output

Set # of decimals for coefficients

Set # of decimals for auxiliary statistics

Set # of decimals for the R²

Set # of decimals for added statistics (addstat option)

Levels of significance

Include some additional statistic, in this case adj. R-sqr. You can select any statistics on the return lists (e-class, r-class or s-class). After running the regression type ereturn list for a list of available statistics.

Type `help outreg2` for more details. If you do not see `outreg2`, you may have to install it by typing `ssc install outreg2`. If this does not work type `findit outreg2`, select from the list and click "install".

Note: If you get the following error message (when you use the option `append` or `replace` it means that you need to close the excel/word window.

```
file results.rtf is read-only; cannot be modified or erased
```

*See the following document for some additional info/tips <http://www.fiu.edu/~tardanic/brianne.pdf>

Regression: publishing regression output (outreg2)

This is how the output would like (you will still need to do some additional editing):

In excel

	A	B
1	v1	v2
2	COEFFICIENT	csat
3		
4	percent	-6.52***
5		(0.51)
6	percent2	0.05***
7		(0.01)
8	high	2.99***
9		(0.49)
10	Constant	844.82***
11		(36.63)
12	Observations	51
13	R-squared	0.93
14	Adj. R-squared	0.92
15	Standard errors in parentheses	
16	*** p<0.001, ** p<0.01, * p<0.05	

In word

COEFFICIENT	csat
percent	-6.52*** (0.51)
percent2	0.05*** (0.01)
high	2.99*** (0.49)
Constant	844.82*** (36.63)
Observations	51
R-squared	0.93
Adj. R-squared	0.92
Standard errors in parentheses	
*** p<0.001, ** p<0.01, * p<0.05	

You can add more models to compare. Lets say you want to add another model without percent2:

`regress csat percent high`

Now type to export the results to excel (**notice** we add the append option)

`outreg2 using results, bdec(2) tdec(2) rdec(2) addec(2) alpha(0.001, 0.01, 0.05) addstat(Adj. R-squared, e(r2_a)) excel append`

In excel

	A	B	C
1	v1	v2	v3
2		(1)	(2)
3	COEFFICIENT	csat	csat
4			
5	percent	-6.52***	-2.32***
6		(0.51)	(0.16)
7	percent2	0.05***	
8		(0.01)	
9	high	2.99***	2.56**
10		(0.49)	(0.76)
11	Constant	844.82***	831.63***
12		(36.63)	(57.39)
13	Observations	51	51
14	R-squared	0.93	0.81
15	Adj. R-squared	0.92	0.80
16	Standard errors in parentheses		
17	*** p<0.001, ** p<0.01, * p<0.05		

In word

	(1)	(2)
COEFFICIENT	csat	csat
percent	-6.52*** (0.51)	-2.32*** (0.16)
percent2	0.05*** (0.01)	
high	2.99*** (0.49)	2.56** (0.76)
Constant	844.82*** (36.63)	831.63*** (57.39)
Observations	51	51
R-squared	0.93	0.81
Adj. R-squared	0.92	0.80
Standard errors in parentheses		
*** p<0.001, ** p<0.01, * p<0.05		

Regression: interaction between dummies

Interaction terms are needed whenever there is reason to believe that the effect of one independent variable depends on the value of another independent variable. We will explore here the interaction between two dummy (binary) variables. In the example below there could be the case that the effect of student-teacher ratio on test scores may depend on the percent of English learners in the district*.

- Dependent variable (Y) – Average test score, variable `testscr` in dataset.
- Independent variables (X)
 - Binary `hi_str`, where '0' if student-teacher ratio (`str`) is lower than 20, '1' equal to 20 or higher.
 - In Stata, first generate `hi_str = 0` if `str < 20`. Then replace `hi_str = 1` if `str >= 20`.
 - Binary `hi_el`, where '0' if English learners (`el_pct`) is lower than 10%, '1' equal to 10% or higher
 - In Stata, first generate `hi_el = 0` if `el_pct < 10`. Then replace `hi_el = 1` if `el_pct >= 10`.
 - Interaction term `str_el = hi_str * hi_el`. In Stata: generate `str_el = hi_str * hi_el`

We will run the regression

```
regress testscr hi_el hi_str str_el, robust
```

```
. regress testscr hi_el hi_str str_el, robust
```

Linear regression

Number of obs =	420
F(3, 416) =	60.20
Prob > F	= 0.0000
R-squared	= 0.2956
Root MSE	= 16.049

testscr	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
hi_el	-18.16295	2.345952	-7.74	0.000	-22.77435	-13.55155
hi_str	-1.907842	1.932215	-0.99	0.324	-5.705964	1.890279
str_el	-3.494335	3.121226	-1.12	0.264	-9.629677	2.641006
_cons	664.1433	1.388089	478.46	0.000	661.4147	666.8718

The equation is $\text{testscr}_{\text{hat}} = 664.1 - 18.1 \cdot \text{hi_el} - 1.9 \cdot \text{hi_str} - 3.5 \cdot \text{str_el}$

The effect of `hi_str` on the test scores is -1.9 but given the interaction term (and assuming all coefficients are significant), the net effect is $-1.9 - 3.5 \cdot \text{hi_el}$. If `hi_el` is 0 then the effect is -1.9 (which is `hi_str` coefficient), but if `hi_el` is 1 then the effect is $-1.9 - 3.5 = -5.4$. In this case, the effect of student-teacher ratio is more negative in districts where the percent of English learners is higher.

See the next slide for more detailed computations.

*The data used in this section is the "California Test Score" data set (`caschool.dta`) from chapter 6 of the book *Introduction to Econometrics* from Stock and Watson, 2003. Data can be downloaded from http://wps.aw.com/aw_stock_ie_2/50/13016/3332253.cw/index.html. For a detailed discussion please refer to the respective section in the book.

Regression: interaction between dummies (cont.)

You can compute the expected values of test scores given different values of `hi_str` and `hi_e1`. To see the effect of `hi_str` given `hi_e1` type the following right after running the regression in the previous slide.

```
. predict yhat1 if hi_str==0 & hi_e1==0
(option xb assumed; fitted values)
(271 missing values generated)

. predict yhat2 if hi_str==1 & hi_e1==0
(option xb assumed; fitted values)
(341 missing values generated)

. predict yhat3 if hi_str==0 & hi_e1==1
(option xb assumed; fitted values)
(331 missing values generated)

. predict yhat4 if hi_str==1 & hi_e1==1
(option xb assumed; fitted values)
(317 missing values generated)
```

These are different scenarios holding constant `hi_e1` and varying `hi_str`. Below we add some labels

```
. label variable yhat1 "Low str/Low e1"
. label variable yhat2 "High str/Low e1"
. label variable yhat3 "Low str/High e1"
. label variable yhat4 "High str/High e1"
```

We then obtain the average of the estimations for the test scores (for all four scenarios, notice same values for all cases).

```
. summarize yhat1 yhat2 yhat3 yhat4
```

variable	Obs	Mean	Std. Dev.	Min	Max
yhat1	149	664.1433	0	664.1433	664.1433
yhat2	79	662.2355	0	662.2355	662.2355
yhat3	89	645.9803	0	645.9803	645.9803
yhat4	103	640.5782	0	640.5782	640.5782

```
. display 664.1 - 662.2
1.9
. display 645.9 - 640.5
5.4
. display 5.4 - 1.9
3.5
```

Here we estimate the net effect of low/high student-teacher ratio holding constant the percent of English learners. When `hi_e1` is 0 the effect of going from low to high student-teacher ratio goes from a score of 664.2 to 662.2, a difference of 1.9. From a policy perspective you could argue that moving from high str to low str improve test scores by 1.9 in low English learners districts.

When `hi_e1` is 1, the effect of going from low to high student-teacher ratio goes from a score of 645.9 down to 640.5, a decline of 5.4 points (1.9+3.5). From a policy perspective you could say that reducing the str in districts with high percentage of English learners could improve test scores by 5.4 points.

Regression: interaction between a dummy and a continuous variable

Lets explore the same interaction as before but we keep student-teacher ratio continuous and the English learners variable as binary. The question remains the same*.

- Dependent variable (Y) – Average test score, variable `testscr` in dataset.
- Independent variables (X)
 - Continuous `str`, student-teacher ratio.
 - Binary `hi_el`, where '0' if English learners (`el_pct`) is lower than 10%, '1' equal to 10% or higher
 - Interaction term `str_el2 = str * hi_el`. In Stata: `generate str_el2 = str*hi_el`

We will run the regression

```
regress testscr str hi_el str_el2, robust
```

```
. regress testscr str hi_el str_el2, robust
```

Linear regression

Number of obs =	420
F(3, 416) =	63.67
Prob > F =	0.0000
R-squared =	0.3103
Root MSE =	15.88

testscr	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]
str	-.9684601	.5891016	-1.64	0.101	-2.126447 .1895268
hi_el	5.639141	19.51456	0.29	0.773	-32.72029 43.99857
str_el2	-1.276613	.9669194	-1.32	0.187	-3.17727 .6240436
_cons	682.2458	11.86781	57.49	0.000	658.9175 705.5742

The equation is $\text{testscr}_{\text{hat}} = 682.2 - 0.97 \cdot \text{str} + 5.6 \cdot \text{hi_el} - 1.28 \cdot \text{str_el2}$

The effect of `str` on `testscr` will be mediated by `hi_el`.

- If `hi_el` is 0 (low) then the effect of `str` is $682.2 - 0.97 \cdot \text{str}$.
- If `hi_el` is 1 (high) then the effect of `str` is $682.2 - 0.97 \cdot \text{str} + 5.6 - 1.28 \cdot \text{str} = 687.8 - 2.25 \cdot \text{str}$

Notice that how `hi_el` changes both the intercept and the slope of `str`. Reducing `str` by one in low EL districts will increase test scores by 0.97 points, but it will have a higher impact (2.25 points) in high EL districts. The difference between these two effects is 1.28 which is the coefficient of the interaction (Stock and Watson, 2003, p.223).

*The data used in this section is the "California Test Score" data set (`caschool.dta`) from chapter 6 of the book *Introduction to Econometrics* from Stock and Watson, 2003. Data can be downloaded from http://wps.aw.com/aw_stock_ie_2/50/13016/3332253.cw/index.html. For a detailed discussion please refer to the respective section in the book.

Regression: interaction between two continuous variables

Lets keep now both variables continuous. The question remains the same*.

- Dependent variable (Y) – Average test score, variable `testscr` in dataset.
- Independent variables (X)
 - Continuous `str`, student-teacher ratio.
 - Continuous `el_pct`, percent of English learners.
 - Interaction term `str_el3 = str * el_pct`. In Stata: `generate str_el3 = str*el_pct`

We will run the regression

```
regress testscr str el_pct str_el3, robust
```

```
. regress testscr str el_pct str_el3, robust
```

Linear regression

					Number of obs =	420
					F(3, 416) =	155.05
					Prob > F =	0.0000
					R-squared =	0.4264
					Root MSE =	14.482

testscr	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
str	-1.117018	.5875135	-1.90	0.058	-2.271884	.0378468
el_pct	-.6729116	.3741231	-1.80	0.073	-1.408319	.0624958
str_el3	.0011618	.0185357	0.06	0.950	-.0352736	.0375971
_cons	686.3385	11.75935	58.37	0.000	663.2234	709.4537

The equation is $\text{testscr}_{\text{hat}} = 686.3 - 1.12 \cdot \text{str} - 0.67 \cdot \text{el_pct} + 0.0012 \cdot \text{str_el3}$

The effect of the interaction term is very small. Following Stock and Watson (2003, p.229), algebraically the slope of `str` is

$-1.12 + 0.0012 \cdot \text{el_pct}$ (remember that `str_el3` is equal to `str*el_pct`). So:

- If `el_pct` = 10, the slope of `str` is -1.108
- If `el_pct` = 20, the slope of `str` is -1.096. A difference in effect of 0.012 points.

In the continuous case there is an effect but is very small (and not significant). See Stock and Watson, 2003, for further details.

*The data used in this section is the "California Test Score" data set (`caschool.dta`) from chapter 6 of the book *Introduction to Econometrics* from Stock and Watson, 2003. Data can be downloaded from http://wps.aw.com/aw_stock_ie_2/50/13016/3332253.cw/index.html. For a detailed discussion please refer to the respective section in the book.

Frequently used Stata commands

Category	Stata commands
Getting on-line help	help search
Operating-system interface	pwd cd sysdir mkdir dir / ls erase copy type
Using and saving data from disk	use clear save append merge compress
Inputting data into Stata	input edit infile infix insheet
The Internet and Updating Stata	update net ado news

Type `help [command name]` in the windows command for details

Source: <http://www.ats.ucla.edu/stat/stata/notes2/commands.htm>

Basic data reporting	describe codebook inspect list browse count assert summarize Table (tab) tabulate
Data manipulation	generate replace egen recode rename drop keep sort encode decode order by reshape
Formatting	format label
Keeping track of your work	log notes
Convenience	display

Is my model OK? (links)

Regression diagnostics: A checklist

<http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter2/statareg2.htm>

Logistic regression diagnostics: A checklist

<http://www.ats.ucla.edu/stat/stata/webbooks/logistic/chapter3/statalog3.htm>

Times series diagnostics: A checklist (pdf)

<http://homepages.nyu.edu/~mrg217/timeseries.pdf>

Times series: dfueller test for unit roots (for R and Stata)

<http://www.econ.uiuc.edu/~econ472/tutorial9.html>

Panel data tests: heteroskedasticity and autocorrelation

- <http://www.stata.com/support/faqs/stat/panel.html>
- <http://www.stata.com/support/faqs/stat/xtreg.html>
- <http://www.stata.com/support/faqs/stat/xt.html>
- http://dss.princeton.edu/online_help/analysis/panel.htm

I can't read the output of my model!!! (links)

Data Analysis: Annotated Output

<http://www.ats.ucla.edu/stat/AnnotatedOutput/default.htm>

Data Analysis Examples

<http://www.ats.ucla.edu/stat/dae/>

Regression with Stata

<http://www.ats.ucla.edu/STAT/stata/webbooks/reg/default.htm>

Regression

<http://www.ats.ucla.edu/stat/stata/topics/regression.htm>

How to interpret dummy variables in a regression

<http://www.ats.ucla.edu/stat/Stata/webbooks/reg/chapter3/statareg3.htm>

How to create dummies

<http://www.stata.com/support/faqs/data/dummy.html>

<http://www.ats.ucla.edu/stat/stata/faq/dummy.htm>

Logit output: what are the odds ratios?

http://www.ats.ucla.edu/stat/stata/library/odds_ratio_logistic.htm

Topics in Statistics (links)

What statistical analysis should I use?

http://www.ats.ucla.edu/stat/mult_pkg/whatstat/default.htm

Statnotes: Topics in Multivariate Analysis, by G. David Garson

<http://www2.chass.ncsu.edu/garson/pa765/statnote.htm>

Elementary Concepts in Statistics

<http://www.statsoft.com/textbook/stathome.html>

Introductory Statistics: Concepts, Models, and Applications

<http://www.psychstat.missouristate.edu/introbook/sbk00.htm>

Statistical Data Analysis

<http://math.nicholls.edu/badie/statdataanalysis.html>

Stata Library. Graph Examples (some may not work with STATA 10)

<http://www.ats.ucla.edu/STAT/stata/library/GraphExamples/default.htm>

Comparing Group Means: The T-test and One-way ANOVA Using STATA, SAS, and SPSS

<http://www.indiana.edu/~statmath/stat/all/ttest/>

Useful links / Recommended books

- DSS Online Training Section <http://dss.princeton.edu/training/>
- UCLA Resources to learn and use STATA <http://www.ats.ucla.edu/stat/stata/>
- DSS help-sheets for STATA http://dss/online_help/stats_packages/stata/stata.htm
- *Introduction to Stata* (PDF), Christopher F. Baum, Boston College, USA. “A 67-page description of Stata, its key features and benefits, and other useful information.” <http://fmwww.bc.edu/GStat/docs/StataIntro.pdf>
- STATA FAQ website <http://stata.com/support/faqs/>

Books

- *Introduction to econometrics* / James H. Stock, Mark W. Watson. 2nd ed., Boston: Pearson Addison Wesley, 2007.
- *Data analysis using regression and multilevel/hierarchical models* / Andrew Gelman, Jennifer Hill. Cambridge ; New York : Cambridge University Press, 2007.
- *Econometric analysis* / William H. Greene. 6th ed., Upper Saddle River, N.J. : Prentice Hall, 2008.
- *Designing Social Inquiry: Scientific Inference in Qualitative Research* / Gary King, Robert O. Keohane, Sidney Verba, Princeton University Press, 1994.
- *Unifying Political Methodology: The Likelihood Theory of Statistical Inference* / Gary King, Cambridge University Press, 1989
- *Statistical Analysis: an interdisciplinary introduction to univariate & multivariate methods* / Sam Kachigan, New York : Radius Press, c1986
- *Statistics with Stata (updated for version 9)* / Lawrence Hamilton, Thomson Books/Cole, 2006