

## Using SPSS for OLS Regression

**Introduction.** This handout summarizes most of the points we cover in Stats I about using SPSS for OLS regression, along with a few additional points. It assumes understanding of the statistical concepts that are presented.

With SPSS, you can get a great deal of information with a single command by specifying various options. This can be quite convenient. However, one consequence of this is that the syntax can get quite complicated. A further complication is that both syntax and features can differ greatly between commands. This probably reflects the way SPSS has evolved over more than 30 years. Stata's syntax and features are, in my opinion, much more logically consistent. Luckily, SPSS's menu structure makes it easy to construct most commands, although some hand-editing may still be necessary; and, for some commands, it may be quicker just to enter the syntax by hand.

**Get the data.** First, open the previously saved data set. (If you prefer, you can also enter the data directly into the program, at least if the data set is not too large.)

```
GET FILE='D:\SOC592\SpssFiles\reg01.sav'.
```

**The Regression Command: Descriptive Statistics, Confidence Intervals, Standardized and Unstandardized Coefficients, VIF and Tolerances, Partial and Semipartial Correlations.** Here is an example regression command with several optional parameters.

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF CI R ANOVA TOL ZPP
  /DEPENDENT income
  /METHOD=ENTER educ jobexp race .
```

Breaking down each part of the command,

<code>/DESCRIPTIVES MEAN STDDEV CORR SIG N</code>	Descriptive statistics. Causes the means, standard deviations, correlation matrix, significance of each correlation, and the sample size to be printed
<code>/MISSING LISTWISE</code>	Listwise deletion of missing data. This means that, if a case is missing data on any of the variables included on the regression command, it will be dropped from the analysis. There are other ways of handling missing data that we will discuss later.
<code>/STATISTICS COEFF CI R ANOVA TOL ZPP</code>	Prints out the unstandardized and standardized coefficients and their T values and significance (COEFF); the 95% confidence interval (CI); Multiple R, $R^2$ , adjusted $R^2$ and standard error of the estimate (R); the Anova Table (ANOVA); the tolerances and variance inflation factors (TOL); and the Zero-order (aka bivariate), Partial, and Part (aka semipartial) correlations of each X with Y (ZPP).
<code>/DEPENDENT income</code>	Specifies the dependent variable. Only one DV can be specified in a single regression command.
<code>/METHOD=ENTER educ jobexp race .</code>	Specifies the block of variables to be included as IVs. In this case, all three variables will be included immediately. Other options are explained below.

Here are excerpts from the output.

## Regression

### Descriptive Statistics

	Mean	Std. Deviation	N
INCOME	24.4150	9.78835	20
EDUC	12.0500	4.47772	20
JOBEXP	12.6500	5.46062	20
RACE	.5000	.51299	20

### Correlations

		INCOME	EDUC	JOBEXP	RACE
Pearson Correlation	INCOME	1.000	.846	.268	-.568
	EDUC	.846	1.000	-.107	-.745
	JOBEXP	.268	-.107	1.000	.216
	RACE	-.568	-.745	.216	1.000
Sig. (1-tailed)	INCOME	.	.000	.127	.005
	EDUC	.000	.	.327	.000
	JOBEXP	.127	.327	.	.180
	RACE	.005	.000	.180	.
N	INCOME	20	20	20	20
	EDUC	20	20	20	20
	JOBEXP	20	20	20	20
	RACE	20	20	20	20

### Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.919 <sup>a</sup>	.845	.816	4.19453

a. Predictors: (Constant), RACE, JOBEXP, EDUC

### ANOVA<sup>b</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1538.920	3	512.973	29.156	.000 <sup>a</sup>
	Residual	281.505	16	17.594		
	Total	1820.425	19			

a. Predictors: (Constant), RACE, JOBEXP, EDUC

b. Dependent Variable: INCOME

### Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Correlations			Collinearity Statistics		
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF	
1	(Constant)	-7.864	5.369		-1.465	.162	-19.246	3.518						
	EDUC	1.981	.323	.906	6.132	.000	1.296	2.666	.846	.838	.603	.442	2.260	
	JOBEXP	.642	.181	.358	3.545	.003	.258	1.026	.268	.663	.348	.947	1.056	
	RACE	.571	2.872	.030	.199	.845	-5.517	6.659	-.568	.050	.020	.427	2.344	

a. Dependent Variable: INCOME

**Hypothesis Testing.** There are a couple of ways to test whether a subset of the variables in a model have zero effects, e.g.  $\beta_1 = \beta_2 = 0$ . One way is to specify a sequence of models and then include the CHA ( $R^2$  change) option, which will present the results of an incremental F test. For example, if we wanted to test whether the effects of EDUC and JOBEXP both equal zero, the syntax would be

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF CI R ANOVA TOL ZPP CHA
  /DEPENDENT income
  /METHOD=ENTER race/ ENTER educ jobexp .
```

With this command, we first estimate a model with RACE only, and then estimate a second model that adds EDUC and JOBEXP. The R Square change info from the following part of the printout tells us whether any of the effects of the variables added in Model 2 significantly differ from 0.

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.568 <sup>a</sup>	.322	.284	8.27976	.322	8.554	1	18	.009
2	.919 <sup>b</sup>	.845	.816	4.19453	.523	27.068	2	16	.000

a. Predictors: (Constant), RACE

b. Predictors: (Constant), RACE, JOBEXP, EDUC

Another approach, and a somewhat more flexible one, is to use METHOD=TEST. With this option, all variables specified by test are entered into the equation. The variable subsets specified by TEST are then deleted and their significance is tested. Multiple subsets can be specified. A sample syntax is

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF CI R ANOVA TOL ZPP CHA
  /DEPENDENT income
  /METHOD=TEST (race) (educ) (jobexp) (race educ) (race jobexp)
  (educ jobexp) (race educ jobexp).
```

As specified here, INCOME will be regressed on RACE, EDUC, and JOBEXP. You'll then get the F values when the vars are dropped one at a time (the F's will equal the corresponding T values squared), two at a time, and then for when all three are dropped (which will be the same as the global F value.) The key part of the printout is

ANOVA<sup>c</sup>

Model			Sum of Squares	df	Mean Square	F	Sig.	R Square Change
1	Subset	RACE	.695	1	.695	.040	.845 <sup>a</sup>	.000
	Tests	EDUC	661.469	1	661.469	37.596	.000 <sup>a</sup>	.363
		JOBEXP	221.054	1	221.054	12.564	.003 <sup>a</sup>	.121
		RACE, EDUC	1408.425	2	704.212	40.026	.000 <sup>a</sup>	.774
		RACE, JOBEXP	236.867	2	118.433	6.731	.008 <sup>a</sup>	.130
		EDUC, JOBEXP	952.476	2	476.238	27.068	.000 <sup>a</sup>	.523
		RACE, EDUC, JOBEXP	1538.920	3	512.973	29.156	.000 <sup>a</sup>	.845
	Regression		1538.920	3	512.973	29.156	.000 <sup>b</sup>	
	Residual		281.505	16	17.594			
	Total		1820.425	19				

a. Tested against the full model.

b. Predictors in the Full Model: (Constant), JOBEXP, EDUC, RACE.

c. Dependent Variable: INCOME

Unfortunately, unlike Stata, SPSS does not provide a convenient way to test hypotheses like  $\beta_1 = \beta_2$ , e.g. the effects of education and job experience are equal. As we will see, however, it is possible to set up your models in such a way that either incremental F tests or (sometimes) T tests can be computed for testing such hypotheses.

**Stepwise Regression.** It is easy to do forward and backwards stepwise regression in SPSS. For example

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF CI R ANOVA TOL ZPP CHA OUTS
  /CRITERIA=PIN(.05) POUT(.10)
  /DEPENDENT income
  /METHOD=FORWARD educ jobexp race .
```

METHOD=FORWARD tells SPSS to do forward stepwise regression; start with no variables and then add them in order of significance. Use METHOD=BACKWARD for backwards selection. The CRITERIA option tells how significant the variable must be to enter into the equation in forward selection (PIN) and how significant it must be to avoid removal in backwards selection (POUT). The OUTS parameter prints statistics about variables not currently in the model, e.g. what their T value would be if the variable was added to the model. SPSS will print detailed information about each intermediate model, whereas Stata pretty much just jumps to the final model. Key parts of the printout include

**Variables Entered/Removed<sup>a</sup>**

Model	Variables Entered	Variables Removed	Method
1	EDUC	.	Forward (Criterion: Probability-of-F-to-enter <= .050)
2	JOBEXP	.	Forward (Criterion: Probability-of-F-to-enter <= .050)

a. Dependent Variable: INCOME

This tells you EDUC got entered first, followed by JOBEXP. RACE did not meet the criteria for entry so it was not included.

**Excluded Variables<sup>a</sup>**

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	JOBEXP	.362 <sup>a</sup>	3.772	.002	.675	.989	1.012	.989
	RACE	.140 <sup>a</sup>	.731	.475	.175	.445	2.245	.445
2	RACE	.030 <sup>b</sup>	.199	.845	.050	.427	2.344	.427

a. Predictors in the Model: (Constant), EDUC

b. Predictors in the Model: (Constant), EDUC, JOBEXP

c. Dependent Variable: INCOME

After model 1, only EDUC is included in the equation. The above tells you that, if JOBEXP were added next, its T value would be 3.772. If instead RACE were added next, its T value would be .731. JOBEXP has the largest T value and is statistically significant, so it gets added in Model 2. If RACE were then added to Model 2, its T value would only be .199. This does not meet the criteria for inclusion, so estimation stops. Here are the coefficients:

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	2.137	3.524		.607	.552	-5.266	9.541					
	EDUC	1.849	.275	.846	6.724	.000	1.271	2.426	.846	.846	.846	1.000	1.000
2	(Constant)	-7.097	3.626		-1.957	.067	-14.748	.554					
	EDUC	1.933	.210	.884	9.209	.000	1.490	2.376	.846	.913	.879	.989	1.012
	JOBEXP	.649	.172	.362	3.772	.002	.286	1.013	.268	.675	.360	.989	1.012

a. Dependent Variable: INCOME

**Sample Selection.** Suppose you only want to analyze a subset of the cases, e.g. blacks. In SPSS, you can use the **Select If** or **Filter** commands. For example,

Select If Race = 0.

would select whites and delete blacks (since race = 1 if black, 0 if white). Note, however, that this is a permanent change, i.e. you can't get the deleted cases back unless you re-open the original data set. If you just want to make temporary sample selections, the `Filter` command is better. The `Filter` command causes variables that have zero or missing values on the filter variable to be excluded from the analysis. However, the cases remain in the working data set and become available again when you specify `Filter Off`. For example, to analyze only whites, we could do something like the following:

```
RECODE RACE (1 = 0)(0 = 1) INTO WHITE.
FILTER BY WHITE.
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF CI R ANOVA TOL ZPP
  /DEPENDENT income
  /METHOD=ENTER educ jobexp .
FILTER OFF.
```

The `recode` command creates a variable coded 1 if white, 0 if black. The `Filter` command then selects the cases coded 1 on white. The `Filter Off` command will cause subsequent analyses to again be done on all the cases. Some of the output includes

**Descriptive Statistics**

	Mean	Std. Deviation	N
INCOME	29.8300	8.34706	10
EDUC	15.3000	2.49666	10
JOBEXP	11.5000	5.93015	10

**Coefficients<sup>a</sup>**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Correlations			Collinearity Statistics		
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF	
1	(Constant)	-13.913	7.828		-1.777	.119	-32.422	4.597						
	EDUC	2.460	.527	.736	4.671	.002	1.214	3.705	.842	.870	.706	.921	1.086	
	JOBEXP	.531	.222	.378	2.398	.048	.007	1.056	.585	.671	.362	.921	1.086	

a. Dependent Variable: INCOME

**Separate Models for Groups.** Suppose you wanted to estimate separate models for both blacks and whites. One way to do this is with the `SORT CASES` and `SPLIT FILE` commands. First, you sort the data by race. Then, you tell SPSS to do separate analyses for each category of race. Use the `Split File Off` command to again do analyses for the entire sample at once. For example,

```
Sort Cases by Race.
Split File Separate by Race.
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF CI R ANOVA TOL ZPP
  /DEPENDENT income
  /METHOD=ENTER educ jobexp .
Split File Off.
```

The **Separate** option on **Split File** displays the split file groups as separate tables, e.g. first you'll get all the results for whites, then you will get the results for blacks. The output includes

## Regression

### RACE = .00 White

Coefficients<sup>a,b</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Correlations			Collinearity Statistics		
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF	
1	(Constant)	-13.913	7.828		-1.777	.119	-32.422	4.597						
	EDUC	2.460	.527	.736	4.671	.002	1.214	3.705	.842	.870	.706	.921	1.086	
	JOBEXP	.531	.222	.378	2.398	.048	.007	1.056	.585	.671	.362	.921	1.086	

a. Dependent Variable: INCOME

b. RACE = .00 White

### RACE = 1.00 Black

Coefficients<sup>a,b</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-6.501	6.406		-1.015	.344	-21.649	8.647					
	EDUC	1.788	.454	.774	3.938	.006	.715	2.862	.740	.830	.771	.994	1.006
	JOBEXP	.707	.324	.429	2.185	.065	-.058	1.473	.369	.637	.428	.994	1.006

a. Dependent Variable: INCOME

b. RACE = 1.00 Black

If you instead specify **Split File Layered by Race** or just **Split File by Race** you get the same information but it is displayed differently, i.e. it is “layered”. Which you use is a matter of personal taste. Here is part of the output.

Coefficients<sup>a</sup>

RACE	Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Correlations			Collinearity Statistics	
			B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
.00 White	1	(Constant)	-13.913	7.828		-1.777	.119	-32.422	4.597					
		EDUC	2.460	.527	.736	4.671	.002	1.214	3.705	.842	.870	.706	.921	1.086
		JOBEXP	.531	.222	.378	2.398	.048	.007	1.056	.585	.671	.362	.921	1.086
1.00 Black	1	(Constant)	-6.501	6.406		-1.015	.344	-21.649	8.647					
		EDUC	1.788	.454	.774	3.938	.006	.715	2.862	.740	.830	.771	.994	1.006
		JOBEXP	.707	.324	.429	2.185	.065	-.058	1.473	.369	.637	.428	.994	1.006

a. Dependent Variable: INCOME

**Analyzing Means, Correlations and Standard Deviations in SPSS.** Sometimes you might want to replicate or modify a published analysis. You don't have the original data, but the authors have published their means, correlations and standard deviations. SPSS lets you input and analyze these directly. Here is how we could analyze our hypothetical income data if we only had the means, correlations and standard deviations available to us:

```
MATRIX data / variables = income educ jobexp race/ format = free full/
  contents = mean sd corr /N = 20.
```

```
BEGIN DATA.
24.415 12.050 12.650 .500
9.78835 4.47772 5.46062 .51299
1.00 .846 .268 -.568
.846 1.000 -.107 -.745
.268 -.107 1.000 .216
-.568 -.745 .216 1.000
END DATA.
```

```
REGRESSION Matrix = In(*)
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /STATISTICS COEFF CI R ANOVA TOL ZPP
  /DEPENDENT income
  /METHOD=ENTER educ jobexp race .
```

Needless to say, this syntax is a little tricky, especially since you can't use Menus to enter it. It helps to be working from a template and/or check the SPSS documentation. Be sure to double-check the output to make sure you have entered the means, correlations and standard deviations correctly. While the syntax may be different, the results are almost exactly the same (the small differences that do exist would likely disappear if we used more decimal places with the correlations):

## Regression

### Descriptive Statistics

	Mean	Std. Deviation	N
INCOME	24.415000	9.7883500	20
EDUC	12.050000	4.4777200	20
JOBEXP	12.650000	5.4606200	20
RACE	.500000	.5129900	20

### Correlations

		INCOME	EDUC	JOBEXP	RACE
Pearson Correlation	INCOME	1.000	.846	.268	-.568
	EDUC	.846	1.000	-.107	-.745
	JOBEXP	.268	-.107	1.000	.216
	RACE	-.568	-.745	.216	1.000
Sig. (1-tailed)	INCOME	.	.000	.127	.004
	EDUC	.000	.	.327	.000
	JOBEXP	.127	.327	.	.180
	RACE	.004	.000	.180	.
N	INCOME	20	20	20	20
	EDUC	20	20	20	20
	JOBEXP	20	20	20	20
	RACE	20	20	20	20

### Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.920 <sup>a</sup>	.846	.817	4.1841156

a. Predictors: (Constant), RACE, JOBEXP, EDUC

### ANOVA<sup>b</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1540.315	3	513.438	29.328	.000 <sup>a</sup>
	Residual	280.109	16	17.507		
	Total	1820.424	19			

a. Predictors: (Constant), RACE, JOBEXP, EDUC

b. Dependent Variable: INCOME

### Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Correlations			Collinearity Statistics		
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF	
1	(Constant)	-7.889	5.359		-1.472	.160	-19.249	3.471						
	EDUC	1.982	.322	.907	6.147	.000	1.299	2.666	.846	.838	.603	.442	2.263	
	JOBEXP	.643	.181	.359	3.557	.003	.260	1.026	.268	.665	.349	.947	1.056	
	RACE	.575	2.866	.030	.201	.844	-5.501	6.651	-.568	.050	.020	.426	2.346	

a. Dependent Variable: INCOME

When analyzing means, correlations and standard deviations, it is important to keep in mind that

- Since you do not have the original data, there is only so much you can do. You can't select subsamples or compute new variables. But, you can do correlational and regression analyses where you analyze different sets of variables.
- Even if you have entered the data correctly, you may not be able to perfectly replicate published results. Simple rounding in the published results (e.g. only reporting correlations to 2 or 3 decimal places) can cause slight differences when replicating analyses. More critically, because of missing data, subsample analyses, and other reasons, cases examined are not always the same throughout an analysis, e.g. 10,000 cases might be analyzed in one regression, 9,700 might be analyzed in another, etc. If you get results that are very different from published results, then the cases used to compute the correlations may be very different from the cases analyzed in that portion of the paper. (Either that, or you've entered the data wrong.)

### Other Comments.

- Unlike Stata, SPSS is not too picky about case, e.g. Regression and REGRESSION both work fine.
- There are various other options for the Regression command as well as other routines that are regression-related. We will talk about some of these in Stats II.