Hierarchical Regression Example
Salary and Publications Example (Cohen, Cohen, West, & Aiken, Table 3.2.1)

Hierarchical regression involves entering variables into the regression model on two or more steps. Any number of variables can be entered on any one step (or "block") and any number of steps can be used. This type of analysis seems less common in recent years, but is sometimes used to examine the percentage of additional variance accounted for by a set of variables over and above the variance accounted for by another set of variables. I illustrate with a very simple two-variable model below, with one variable entered on each step, in order to compare to the simultaneous model illustrated in the previous multiple regression example handout. The term "hierarchical" is sometimes used for a different analysis, called "hierarchical linear modeling" designed for nested data like schools, and should not be confused with the regression approach illustrated here. There is also an approach to building regression models called "stepwise regression," which is an exploratory approach (discussed soon) and is not the same as this approach, because in hierarchical regression the researcher chooses which variables to enter in which order.

Below are two hierarchical regression examples, both with the number of publications and the years since PhD predicting salary, but with the two variables entered in different orders. It would be very uncommon to retest a model using different orderings, but I want to illustrate that on the final step, the two models are the same and equal to the previously illustrated simultaneous model results, and the order that they are entered does not matter once you get to the final step. See also that the first step with just one variable is the same as a simple regression or correlation.

Entering PUBS First

SPSS Syntax

```
correlations vars=pubs time salary.

regression vars=salary time pubs
    /descriptives=mean stdev
    /statistics=anova r coeff ses cha
    /dependent=salary
    /method=enter pubs /enter time.
```

Correlations

		pubs number of publications	time years since PhD	salary annual salary in dollars
pubs number of publications	Pearson Correlation	1	.657	.588
	Sig. (2-tailed)		.008	.021
	N	15	15	15
time years since PhD	Pearson Correlation	.657	1	.710
	Sig. (2-tailed)	.008		.003
	N	15	15	15
salary annual salary in dollars	Pearson Correlation	.588	.710	1
	Sig. (2-tailed)	.021	.003	
	N	15	15	15

Model Summary

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.588 [a]	.346	.295	6623.61975	.346	6.864	1	13	.021
2	.728 [b]	.530	.452	5839.23054	.185	4.727	1	12	.050

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	301137778.67	1	301137778.67	6.864	.021 [b]
	Residual	570340400.93	13	43872338.533		
	Total	871478179.60	14			
2	Regression	462318820.34	2	231159410.17	6.780	.011 [c]
	Residual	409159359.26	12	34096613.272		
	Total	871478179.60	14			

a. Dependent Variable: salary annual salary in dollars

b. Predictors: (Constant), pubs number of publications

c. Predictors: (Constant), pubs number of publications, time years since PhD

Coefficients [a]

		Unstandardized Coefficients		Standardized Coefficients			
Model		B	Std. Error	Beta	Std. Error	t	Sig.
1	(Constant)	46357.449	3072.730			15.087	<.001
	pubs number of publications	335.526	128.067	.588	.224	2.620	.021
2	(Constant)	43082.394	3099.493			13.900	<.001
	pubs number of publications	121.801	149.699	.213	.262	.814	.432
	time years since PhD	982.867	452.057	.570	.262	2.174	.050

a. Dependent Variable: salary annual salary in dollars

Entering TIME First

Syntax

```
regression vars=salary time pubs
    /descriptives=mean stdev
    /statistics=anova r coeff ses cha
    /dependent=salary
    /method=enter time /enter pubs.
```

Variables Entered/Removed [a]

Model	Variables Entered	Variables Removed	Method
1	time years since PhD [b]	.	Enter
2	pubs number of publications [b]	.	Enter

a. Dependent Variable: salary annual salary in dollars

b. All requested variables entered.

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	.710 [a]	.505	.466	5762.82286	.505	13.241	1	13	.003
2	.728 [b]	.530	.452	5839.23054	.026	.662	1	12	.432

a. Predictors: (Constant), time years since PhD

b. Predictors: (Constant), time years since PhD, pubs number of publications

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	439746525.09	1	439746525.09	13.241	.003 [b]
	Residual	431731654.51	13	33210127.270		
	Total	871478179.60	14			
2	Regression	462318820.34	2	231159410.17	6.780	.011 [c]
	Residual	409159359.26	12	34096613.272		
	Total	871478179.60	14			

a. Dependent Variable: salary annual salary in dollars

b. Predictors: (Constant), time years since PhD

c. Predictors: (Constant), time years since PhD, pubs number of publications

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients		t	Sig.
		B	Std. Error	Beta	Std. Error		
1	(Constant)	43658.594	2978.022			14.660	<.001
	time years since PhD	1224.392	336.476	.710	.195	3.639	.003
2	(Constant)	43082.394	3099.493			13.900	<.001
	time years since PhD	982.867	452.057	.570	.262	2.174	.050
	pubs number of publications	121.801	149.699	.213	.262	.814	.432

a. Dependent Variable: salary annual salary in dollars

R Code (Pubs-First Model Only)

For the hierarchical model in R, I use the `lm` function and test two separate models then request the comparison using the anova function. Only the model with publications entered first is illustrated.

```
> #hierarchical regression using base R function lm
> model1 = lm(SALARY~PUBS, data=d)
> model2 = lm(SALARY~PUBS + TIME, data=d)
> summary(model1)

Call:
lm(formula = SALARY ~ PUBS, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-10597.7  -6320.8   -520.9   5328.4   9801.6

Coefficients:
            Estimate Std. Error t value      Pr(>|t|)
(Intercept)  46357.4     3072.7   15.09 0.00000000129
PUBS           335.5      128.1    2.62        0.0212

Residual standard error: 6624 on 13 degrees of freedom
Multiple R-squared:  0.3455, Adjusted R-squared:  0.2952
F-statistic: 6.864 on 1 and 13 DF,  p-value: 0.02119

> summary(model2)

Call:
lm(formula = SALARY ~ PUBS + TIME, data = mydata)

Residuals:
   Min     1Q Median     3Q    Max
-12066  -3522   -342   3324   8847

Coefficients:
            Estimate Std. Error t value      Pr(>|t|)
(Intercept)  43082.4     3099.5  13.900 0.00000000926
PUBS           121.8      149.7   0.814        0.4317
TIME           982.9      452.1   2.174        0.0504

Residual standard error: 5839 on 12 degrees of freedom
Multiple R-squared:  0.5305, Adjusted R-squared:  0.4522
F-statistic:  6.78 on 2 and 12 DF,  p-value: 0.01071

> anova(model1,model2)
Analysis of Variance Table

Model 1: SALARY ~ PUBS
Model 2: SALARY ~ PUBS + TIME
  Res.Df       RSS Df Sum of Sq      F  Pr(>F)
1     13 570340401
2     12 409159359  1 161181042 4.7272 0.05041

#use the z-score method or the lm.beta package as previously illustrated for standardized coefficients for
either the first or second model.
```

Write-up

I won't provide an additional write-up here, because it would be mostly redundant with the write-up for the simultaneous regression. To report a hierarchical regression, be sure to state that a hierarchical approach was used, which variables were entered on which step, and include the R-squared change and significance (e.g., for the second step, R^2 change = .03, $F(1,12) = .66$, p = .43) for each group of variables added to the model (except the first block). A table will likely be helpful if there are several variables added at a time and/or several steps. See "Overhead: Good Example of Hierarchical Table" for an illustration.