O.P. JINDAL GLOBAL UNIVERSITY, SONIPAT, HARYANA

MULTIPLE REGRESSION

An Overview through SPSS

Prof (Dr.) Krishan K. Pandey

2011

WEB PAGE: KRISHANPANDEY.COM, EMAIL:KKPANDEY@JGU.EDU.IN

Multiple Regression

Assumptions of Multiple Linear Regression

There are many assumptions to consider but we will focus on the major ones that are easily tested with SPSS. The assumptions for multiple regression include the following: that the relationship between each of the predictor variables and the dependent variable is linear and that the error, or residual, is normally distributed and uncorrelated with the predictors. A condition that can be extremely problematic as well is **multicollinearity,** which can lead to misleading and/or inaccurate results. Multicollinearity (or collinearity) occurs when there are high intercorrelations among some set of the predictor variables. In other words, multicollinearity happens when two or more predictors contain much of the same information.

Although a correlation matrix indicating the intercorrelations among all pairs of predictors is helpful in determining whether multicollinearity is a problem, it will not always indicate that the condition exists.

Multicollinearity may occur because several predictors, taken *together*, are related to some other predictors or set of predictors. For this reason, it is important to test for multicollinearity when doing multiple regression.

There are several different ways of computing multiple regression that are used under somewhat different circumstances. We will have you use several of these approaches, so that you will be able to see that the method one uses to compute multiple regression influences the information one obtains from the analysis. If the researcher has no prior ideas about which variables will create the best prediction equation and has a reasonably small set of predictors, then **simultaneous regression** is the best method to use. It is preferable to use the **hierarchical method** when one has an idea about the *order* in which one wants to enter predictors and wants to know how prediction by certain variables *improves on* prediction by others. Hierarchical regression appropriately corrects for capitalization on chance; whereas, **stepwise,** another method available in SPSS in which variables are entered sequentially, does not. Both simultaneous regression and hierarchical regression require that you specify exactly which variables serve as predictors. Sometimes you have a relatively large set of variables that may be good predictors of the dependent variable, but you cannot enter such a large set of variables without sacrificing the power to find significant results. In such a case, stepwise regression might be used. However, as indicated earlier, stepwise regression capitalizes on chance more than many researchers find acceptable.

• Retrieve your data file: **hsbdataB.sav**

Problem 6.1: Using the Simultaneous Method to Compute Multiple Regression

To reiterate, the purpose of multiple regression is to predict an interval (or scale) dependent variable from a combination of several interval/scale, and/or dichotomous independent/predictor variables. In the following assignment, we will see *if math achievement* can be predicted better from a combination of several of our other variables, such as the *motivation scale, grades in high school,* and *mother's* and *father's education.* In Problems 6.1 and 6.3, we will run the multiple regression using alternate methods provided by SPSS. In Problem 6.1, we will assume that all seven of the predictor variables are important and that we want to see what is the highest possible multiple correlation of these variables with the dependent variable. For this purpose, we will use the method that SPSS calls **Enter** (often called **simultaneous regression),** which tells the computer to consider all the variables at the same time. In Problem 6.3, we will use the hierarchical method. 6.1. How well can you predict *math achievement* from a combination of seven variables: *motivation, competence,pleasure, grades in high school, father's education, mother's education, andgendert*

In this problem, the computer will enter/consider all the variables at the same time. Also, we will ask which of these seven predictors contribute significantly to the multiple correlation/regression.

It is a good idea to check the correlations among the predictor variables prior to running the multiple regression, to determine if the predictors are sufficiently correlated such that multicollinearity is highly likely to be a problem. This is especially important to do when one is using a relatively large set of predictors, and/or if, for empirical or conceptual reasons, one believes that some or all of the predictors might be highly correlated. If variables are highly correlated (e.g., correlated at .50 or .60 and above), then one might decide to combine (aggregate) them into a composite variable or eliminate one or more of the highly correlated variables if the variables do not make a meaningful composite variable. For this example, we will check correlations between the variables to see if there might be multicollinearity problems. We typically also would create a scatterplot matrix to check the assumption of linear relationships of each predictor with the dependent variable and a scatterplot between the predictive equation and the residual to check for the assumption that these are uncorrelated. In this problem, we will not do so because we will show you how to do these assumption checks in Problem 6.2.

• Click on **Analyze => Correlate => Bivariate. The Bivariate Correlations** window will appear.

• Select the variables *motivation scale, competence scale, pleasure scale, grades in h.s., father's education, mother's education,* and *gender* and click them over to the **Variables** box.

• Click on **Options => Missing values => Exclude cases listwise.**

• Click on Continue and then click on OK. A correlation matrix like the one in Output 6.la should appear.

Output 6.1a: Correlation Matrix

```
CORRELATIONS
  /VARIABLES=motivation competence pleasure grades faed maed gender
  /PRINT=TWOTAIL NOSIG
  /MISSING=LISTWISE.
```

Correlations

High correlations among predictors indicate it is likely that there will be a problem with multicollinearity.

Correlations[a]

		motivation scale	competence scale	pleasure scale	grades in h.s.	father's education	mother's education	gender
motivation scale	Pearson Correlation	1	.517**	.277*	.020	.049	.115	-.178
	Sig. (2-tailed)	.	.000	.021	.872	.692	.347	.143
competence scale	Pearson Correlation	.517**	1	.413**	.216	.031	.234	-.037
	Sig. (2-tailed)	.000	.	.000	.075	.799	.053	.760
pleasure scale	Pearson Correlation	.277*	.413**	1	-.081	.020	.108	.084
	Sig. (2-tailed)	.021	.000	.	.509	.869	.378	.492
grades in h.s.	Pearson Correlation	.020	.216	-.081	1	.315**	.246*	.162
	Sig. (2-tailed)	.872	.075	.509	.	.008	.042	.182
father's education	Pearson Correlation	.049	.031	.020	.315**	1	.649**	-.266*
	Sig. (2-tailed)	.692	.799	.869	.008	.	.000	.027
mother's education	Pearson Correlation	.115	.234	.108	.246*	.649**	1	-.223
	Sig. (2-tailed)	.347	.053	.378	.042	.000	.	.065
gender	Pearson Correlation	-.178	-.037	.084	.162	-.266*	-.223	1
	Sig. (2-tailed)	.143	.760	.492	.182	.027	.065	.

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

a. Listwise N=69

The correlation matrix indicates large correlations between *motivation* and *competence* and between *mother's education and father's education.* To deal with this problem, we would usually aggregate or eliminate variables that are highly correlated. However, we want to show how the collinearity problems created by these highly correlated predictors affect the Tolerance values and the significance of the beta coefficients, so we will run the regression without altering the variables. To run the regression, follow the steps below:

• Click on the following: **Analyze => Regression => Linear. The Linear Regression** window (Fig. 6.1) should appear.

• Select *math achievement* and click it over to the **Dependent** box (dependent variable).

• Next select the variables *motivation scale, competence scale, pleasure scale, grades in h.s., father's education, mother's education,* and *gender* and click them over to the **Independent(s)** box (independent variables).

• Under **Method,** be sure that **Enter** is selected.

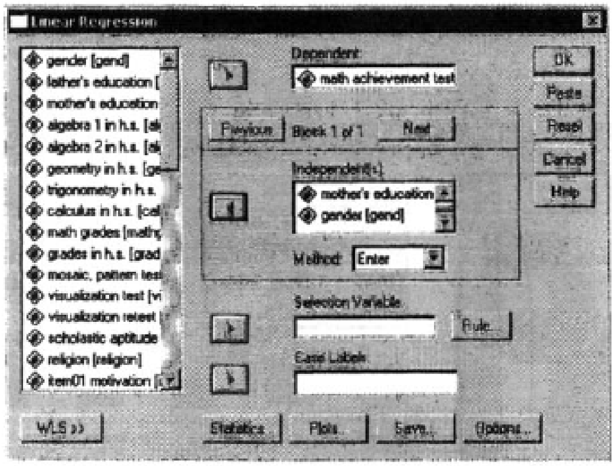


Fig. 6.1. Linear Regression.

- Click on **Statistics,** click on **Estimates** (under **Regression Coefficients**), and click on **Model fit, Descriptives,** and **Collinearity diagnostics**. (See Fig. 6.2.)

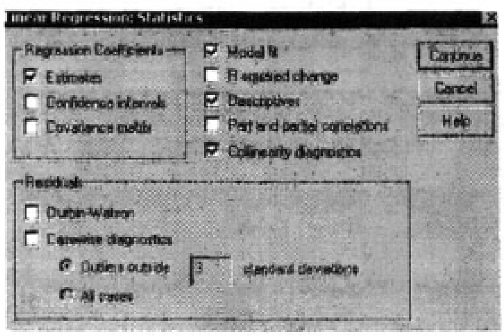


Fig. 6.2. Linear Regression: Statistics.

- Click on **Continue.**
- Click on **OK.**

Compare your output and syntax to Output 6.1b.

Output 6.1b: Multiple Linear Regression, Method = Enter

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA COLLIN TOL
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT mathach
  /METHOD=ENTER motivation competence pleasure grades faed maed gender.
```

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
math achievement test	12.7536	6.66293	69
Motivation scale	2.8913	.62676	69
Competence scale	3.3188	.62262	69
pleasure scale	3.1667	.66789	69
grades in h.s.	5.71	1.573	69
father's education	4.65	2.764	69
mother's education	4.07	2.185	69
gender	.54	.502	69

N is 69 because 6 participants have some missing data.

This is a repeat of the correlation matrix we did earlier, indicating high correlations among predictors.

Correlations with *math achievement.*

Correlations

		math achievement test	Motivation scale	Competence scale	pleasure scale	grades in h.s.	father's education	mother's education	gender
Pearson Correlation	math achievement test	1.000	.256	.260	.087	.470	.416	.387	-.272
	Motivation scale	.256	1.000	.517	.274	.020	.049	.115	-.178
	Competence scale	.260	.517	1.000	.438	.216	.031	.234	-.037
	pleasure scale	.087	.274	.438	1.000	-.111	-.008	.085	.037
	grades in h.s.	.470	.020	.216	-.111	1.000	.315	.246	.162
	father's education	.416	.049	.031	-.008	.315	1.000	.649	-.266
	mother's education	.387	.115	.234	.085	.246	.649	1.000	-.223
	gender	-.272	-.178	-.037	.037	.162	-.266	-.223	1.000
Sig. (1-tailed)	math achievement test	.	.017	.015	.239	.000	.000	.001	.012
	Motivation scale	.017	.	.000	.011	.436	.346	.173	.072
	Competence scale	.015	.000	.	.000	.037	.400	.026	.380
	pleasure scale	.239	.011	.000	.	.182	.474	.244	.383
	grades in h.s.	.000	.436	.037	.182	.	.004	.021	.091
	father's education	.000	.346	.400	.474	.004	.	.000	.014
	mother's education	.001	.173	.026	.244	.021	.000	.	.032
	gender	.012	.072	.380	.383	.091	.014	.032	.
N	math achievement test	69	69	69	69	69	69	69	69
	Motivation scale	69	69	69	69	69	69	69	69
	Competence scale	69	69	69	69	69	69	69	69
	pleasure scale	69	69	69	69	69	69	69	69
	grades in h.s.	69	69	69	69	69	69	69	69
	father's education	69	69	69	69	69	69	69	69
	mother's education	69	69	69	69	69	69	69	69
	gender	69	69	69	69	69	69	69	69

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	gender, pleasure scale, grades in h.s., Motivation scale, mother's education, Competence scale, father's education[a]		Enter

a. All requested variables entered.

b. Dependent Variable: math achievement test

Significance level of correlations with *math achievement.*

Multiple correlation coefficient.

Indicates that 36% of the variance can be predicted from the independent variables.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.654[a]	.427	.362	5.32327

a. Predictors: (Constant), gender, pleasure scale, grades in h.s., Motivation scale, mother's education, Competence scale, father's education

b. Dependent Variable: math achievement test

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1290.267	7	184.324	6.505	.000[a]
	Residual	1728.571	61	28.337		
	Total	3018.838	68			

a. Predictors: (Constant), gender, pleasure scale, grades in h.s., Motivation scale, mother's education, Competence scale, father's education

b. Dependent Variable: math achievement test

Indicates that the combination of these variables significantly ($p < .001$) predicts the dependent variable.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-6.912	4.749		-1.455	.151		
	Motivation scale	1.639	1.233	.154	1.330	.188	.698	1.432
	Competence scale	1.424E-02	1.412	.001	.010	.992	.539	1.854
	pleasure scale	.953	1.119	.096	.852	.398	.746	1.340
	grades in h.s.	1.921	.480	.453	4.001	.000	.731	1.368
	father's education	.303	.331	.126	.915	.364	.497	2.013
	mother's education	.333	.406	.109	.820	.415	.529	1.892
	gender	-3.497	1.424	-.264	-2.455	.017	.814	1.228

a. Dependent Variable: math achievement test

Only *grades* and *gender* are significantly contributing to the equation. However, all of the variables need to be included to obtain this result, since the overall F value was computed with all the variables in the equation.

Tolerance and VIF give the same information. (Tolerance = 1/VIF) They tell us if there is multicollinearity. If the Tolerance value is low ($< 1-R^2$), then there is probably a problem with multicollinearity. In this case, since adjusted R^2 is .36, and $1- R^2$ is about .64, then tolerances are low for *competence*, *mother's* and *father's*

This tells you how much each variable is contributing to any collinearity in the model.

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions							
				(Constant)	Motivation scale	Competence scale	pleasure scale	grades in h.s.	father's education	mother's education	gender
1	1	7.035	1.000	.00	.00	.00	.00	.00	.00	.00	.00
	2	.550	3.577	.00	.00	.00	.00	.00	.04	.02	.49
	3	.215	5.722	.00	.02	.01	.01	.00	.18	.09	.32
	4	8.635E-02	9.026	.00	.00	.00	.00	.06	.45	.78	.01
	5	5.650E-02	11.159	.00	.01	.00	.10	.60	.23	.04	.08
	6	2.911E-02	15.545	.01	.59	.00	.43	.05	.01	.00	.08
	7	1.528E-02	21.456	.70	.00	.46	.02	.02	.05	.04	.01
	8	1.290E-02	23.350	.29	.38	.53	.44	.28	.03	.02	.00

a. Dependent Variable: math achievement test

Interpretation of Output 6.1

First, the output provides the usual descriptive statistics for all eight variables. Note that the *N* is 69 because 6 participants are missing a score on one or more variables. Multiple regression uses only the participants who have complete data for all the variables. The next table is a correlation matrix similar to the one in Output 6.1a. Note that the first column shows the correlations of the other variables with *math achievement* and that *motivation*, *competence*, *grades in high school*, *father's education*, *mother's education*, and *gender* are all significantly correlated with *math achievement*. As we observed before, several of the predictor/ independent variables are highly correlated with each other; that is, *competence* and *motivation* (.517) and *mother's education* and *father's education* (.649).

The **Model Summary** table shows that the multiple correlation coefficient (***R***), using all the predictors simultaneously, is .65 ($R^2 = .43$) and the **adjusted** R^2 is .36, meaning that 36% of the variance in *math achievement* can be predicted from *gender*, *competence*, etc. combined. Note that the adjusted R^2 is lower than the unadjusted R^2. This is, in part, related to the number of variables in the equation. The adjustment is also affected by the magnitude of the effect and the sample size. As you will see from the coefficients table, only *father's education* and *gender* are significant, but the other five variables will always add a little to the prediction of *math achievement*. Because so many independent variables were used, a reduction in the number of variables might help us find an equation that explains more of the variance in the dependent variable. It is helpful to use the concept of parsimony with multiple regression, and use the smallest number of predictors needed.

The ANOVA table shows that ***F* = 6.51** and is significant. This indicates that the combination of the predictors significantly predict *math achievement*.

One of the most important tables is the **Coefficients** table. It indicates the **standardized beta coefficients**, which are interpreted similarly to correlation coefficients or factor weights (see chapter 4). The ***t*** value and the **Sig** opposite each independent variable indicates whether that variable is significantly contributing to the equation for predicting *math achievement* from the whole set of predictors. Thus, *h.s. grades* and *gender*, in this example, are the only variables that are significantly adding anything to the prediction when the other five variables are already considered. It is important to note that all the variables are being considered together when these values are computed. Therefore, if you delete one of the predictors that is not significant, it can affect the levels of significance for other predictors.

However, as the **Tolerances** in the **Coefficients** table suggest, and as we will see in Problem 6.2, these results are somewhat misleading. Although the two parent education measures were significantly correlated with *math achievement*, they did not contribute to the multiple regression predicting *math achievement*. What has happened here is that these two measures were also highly correlated with each other, and multiple regression eliminates all overlap between predictors. Thus, neither *father's education* nor *mother's education* had much to contribute when the other was also used as a predictor. Note that tolerance for each of these variables is < .64 (1-.36), indicating that too much multicollinearity (overlap between predictors) exists. The same is true for *competence*, once *motivation* is entered. One way to handle multicollinearity is to combine variables that are highly related if that makes conceptual sense. For example, you could make a new variable called *parents' education*, as we will for Problem 6.2.

Problem 6.2: Simultaneous Regression Correcting Multicollinearity

In Problem 6.2, we will use the combined/average of the two variables, *mother's education and father's education, and* then recompute the multiple regression, after omitting *competence and pleasure.*

We *combined father's education* and *mother's education* because it makes conceptual sense and because these two variables are quite highly related *(r* = .65). We know that entering them as two separate variables created problems with multicollinearity because tolerance levels were low for these two variables, and, despite the fact that both variables were significantly and substantially correlated with *math achievement,* neither contributed significantly to predicting *math achievement* when taken together. When it does not make sense to combine the highly correlated variables, one can eliminate one or more of them. Because the conceptual distinction between *motivation, competence, and pleasure* was important for us, and because *motivation* was more important to us than *competence or pleasure,* we decided to delete the latter two scales from the analysis. We wanted to see *if motivation* would contribute to the prediction *of math achievement* if its contribution was not canceled out by *competence* and/or *pleasure. Motivation* and *competence* are so highly correlated that they create problems with multicollinearity. We eliminate *pleasure* as well, even though its tolerance is acceptable, because it is virtually uncorrelated with *math achievement,* the dependent variable, and yet it is correlated with *motivation* and *competence.* Thus, it is unlikely to contribute meaningfully to the prediction *of mathachievement,* and its inclusion would only serve to reduce power and potentially reduce the predictive power *of motivation.* It would be particularly important to eliminate a variable such *as pleasure* if it were strongly correlated with another predictor, as this can lead to particularly misleading results.

6.2. Rerun Problem 6.1 using the *parents' education* variable *(parEduc)* instead *offaed* and *maed* and omitting the *competence and pleasure scales.* First, we created a matrix scatterplot (as in chapter 2) to see if the variables are related to each other in a linear fashion. You can use the syntax in Output 6.2 or use the **Analyze => Scatter** windows as shown below.

• Click on **Graphs => Scatter...**

• Select **Matrix** and click on **Define.**

• Move *math achievement, motivation, grades, parent's education, and gender* into the **Matrix Variables: box.**

• Click on **Options.** Check to be sure that **Exclude cases listwise** is selected.

• Click on **Continue** and then **OK.**

Then, run the regression, using the following steps:

• Click on the following: **Analyze => Regression => Linear. The Linear Regression** window (Fig. 6.1) should appear. This window may still have the variables moved over to the **Dependent** and **Independent(s)** boxes. If so, click on **Reset.**

• Move *math achievement* into the **Dependent box.**

• Next select the variables *motivation, grades in h.s., parent's education, and gender and* move them into the **Independent(s)** box (independent variables).

• Under **Method,** be sure that **Enter** is selected.

• Click on **Statistics,** click on **Estimates** (under **Regression Coefficients),** and click on **Model fit, Descriptives,** and **Collinearity diagnostics** (See Fig. 6.2.).

• Click on **Continue.**

• Click on **OK.**

Then, we added a plot to the multiple regression to see the relationship of the predictors and the residual. To make this plot follow these steps:

• Click on **Plots... (in Fig. 6.1 to** get Fig. **6.3.)**

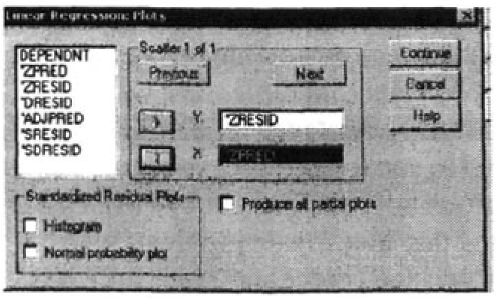


Fig. 6.3. Linear Regression: Plots.

- **Move ZRESID to the Y: box.**
- **Move ZPRED to the X: box. This enables us to check the assumption that the predictors an residual are uncorrelated.**
- Click on **Continue**.
- Click on **OK**.

Refer to Output 6.2 for comparison.

Output 6.2: Multiple Linear Regression with Parent's Education, Method = Enter

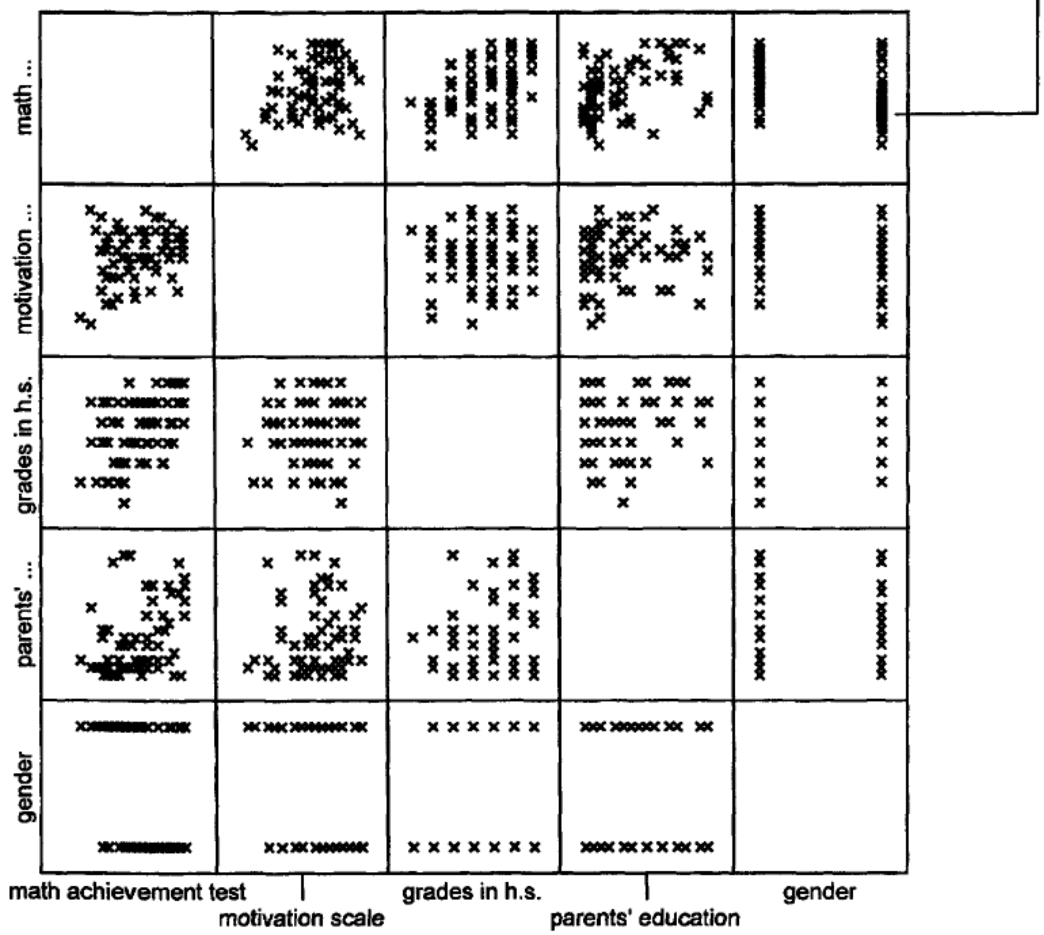
```
GRAPH
  /SCATTERPLOT(MATRIX)=mathach motivation grades parEduc gender
  /MISSING=LISTWISE .


REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA COLLIN TOL
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT mathach
  /METHOD=ENTER motivation grades parEduc gender
  /SCATTERPLOT=(*ZRESID , *ZPRED) .
```

Graph

The top row shows four scatterplots (relationships) of the dependent variables with each of the predictors. To meet the assumption of linearity a straight line, as opposed to a curved line, should fit the points relatively well.

Dichotomous variables have two columns (or rows) of data points. If the data points bunch up near the top of the left column and the bottom of the right, the correlation will be negative (and vice versa). Linearity would be violated if the data points bunch at the center of one column and at the ends of the other column.



Regression

Descriptive Statistics

	Mean	Std. Deviation	N
math achievement test	12.6028	6.75676	73
motivation scale	2.8744	.63815	73
grades in h.s.	5.68	1.589	73
parents' education	4.3836	2.30266	73
gender	.55	.501	73

Note that $N = 73$, indicating that eliminating competence and pleasure reduced the amount of missing data.

Note that all the predictors are significantly related to *math achievement.*

None of the relationships among predictors is greater than .25.

Correlations

		math achievement test	motivation scale	grades in h.s.	parents' education	gender
Pearson Correlati	math achievement t	1.000	.316	.504	.394	-.303
	motivation scale	.316	1.000	.084	.090	-.209
	grades in h.s.	.504	.084	1.000	.250	.115
	parents' education	.394	.090	.250	1.000	-.227
	gender	-.303	-.209	.115	-.227	1.000
Sig. (1-tailed)	math achievement t	.	.003	.000	.000	.005
	motivation scale	.003	.	.241	.225	.038
	grades in h.s.	.000	.241	.	.016	.166
	parents' education	.000	.225	.016	.	.027
	gender	.005	.038	.166	.027	.
N	math achievement t	73	73	73	73	73
	motivation scale	73	73	73	73	73
	grades in h.s.	73	73	73	73	73
	parents' education	73	73	73	73	73
	gender	73	73	73	73	73

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	gender, grades in h.s., motivation scale, parents' education[a]		Enter

a All requested variables entered.
b Dependent Variable: math achievement test

This indicates we used simultaneous regression in this problem.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.678[a]	.459	.427	5.11249

a. Predictors: (Constant), gender, grades in h.s., motivation scale, parent's education

b. Dependent Variable: math achievement test

The Adjusted R Square indicates that we have a fairly good model, explaining about 43% of the variance in *math achievement.*

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1509.723	4	377.431	14.440	.000[a]
	Residual	1777.353	68	26.138		
	Total	3287.076	72			

a. Predictors: (Constant), gender, grades in h.s., motivation scale, parent's education

b. Dependent Variable: math achievement test

Our model significantly predicts *math achievement.*

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-5.444	3.605		-1.510	.136		
	motivation scale	2.148	.972	.203	2.211	.030	.944	1.059
	grades in h.s.	1.991	.400	.468	4.972	.000	.897	1.115
	parent's education	.580	.280	.198	2.070	.042	.871	1.148
	gender	-3.631	1.284	-.269	-2.828	.006	.877	1.141

a. Dependent Variable: math achievement test

Here are the values to check for multicollinearity. Note that all tolerances are well over .57 ($1-R^2$).

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions				
				(Constant)	motivation scale	grades in h.s.	parent's education	gender
1	1	4.337	1.000	.00	.00	.00	.01	.01
	2	.457	3.082	.00	.00	.00	.07	.68
	3	.135	5.665	.02	.07	.02	.85	.17
	4	.052	9.120	.01	.20	.87	.06	.06
	5	.019	15.251	.97	.73	.11	.01	.08

a. Dependent Variable: math achievement test

Casewise Diagnostics[a]

Case Number	Std. Residual	math achievement test
63	-3.174	1.00

a. Dependent Variable: math achievement test

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	1.5029	22.2180	12.6028	4.57912	73
Residual	-16.2254	10.3169	.0000	4.96845	73
Std. Predicted Value	-2.424	2.100	.000	1.000	73
Std. Residual	-3.174	2.018	.000	.972	73

a. Dependent Variable: math achievement test

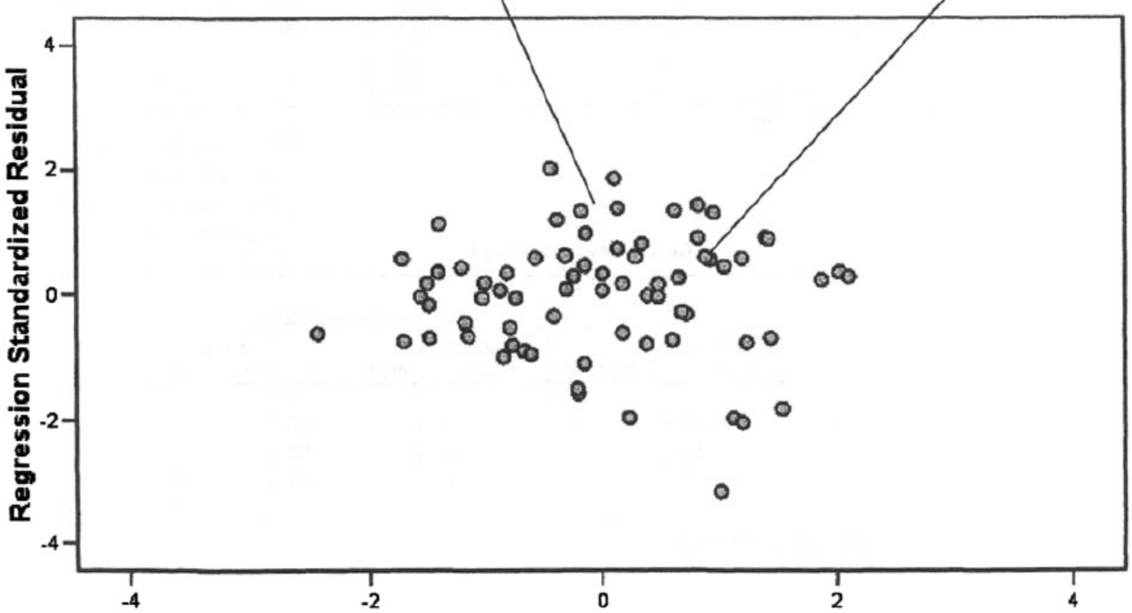
Charts

If the dots created a pattern, this would indicate the residuals are not normally distributed, the residual is correlated with the independent variables, and/or the variances of the residuals are not constant.

Because the dots are scattered, it indicates the data meet the assumptions of the errors being normally distributed and the variances of the residuals being constant.

Scatterplot

Dependent Variable: math achievement test



Interpretation of Output 6.2

This output begins with the **scatterplot matrix**, which shows that the independent variables are generally linearly related to the dependent variable of *math achievement*, meeting this assumption. One should check the matrix scatterplots to see if there are curvilinear relationships between any of the variables; in this example, there are none. If the variables had not met this assumption, we could have transformed them, aggregated some, and/or eliminated some independent variables. (See chapter 2 for how to do transformations.) There are also only low to moderate relationships among the predictor variables in the **Correlations** table. This is good. The other assumptions are checked in the **residual scatterplot** at the end of the output, which indicates that the errors are normally distributed, the variances of the residuals are constant, and the residual is relatively uncorrelated with the linear combination of predictors.

The next important part of the output to check is the **Tolerance** and VIF values in the **Coefficients** table for the existence of multicollinearity. In this example, we do not need to worry about multicollinearity because the Tolerance values are close to 1.

The **Model Summary** table gives the **R** (.68) and **Adjusted R square** (.43). Thus, this model is predicting 43% of the variance in *math achievement*. If we want to compare this model to that in Problem 6.1, we use the Adjusted R square to see which model is explaining more of the variance in the dependent variable. Interestingly, this model is predicting more of the variance in *math achievement* than the model in Problem 6.1, despite using fewer predictors.

As can be seen from the **ANOVA** table, the model of *motivation, grades in h.s., parents' education*, and *gender* significantly predicts *math achievement*, $F(4, 68) = 14.44, p < .001$.

We can see from the **Coefficients** table that now <u>all</u> of the predictors are significantly contributing to the equation (see the **Sig**. column).

How to Write About Output 6.2

Results

Multiple regression was conducted to determine the best linear combination of gender, grades in h.s., parents' education, and motivation for predicting math achievement test scores. The means, standard deviations, and intercorrelations can be found in Table 6.1. This combination of variables significantly predicted math achievement, $F(4,68) = 14.44, p < .001$, with all four variables significantly contributing to the prediction. The beta weights, presented in Table 6.2, suggest that good grades in high school contribute most to predicting math achievement, and that being male, having high math motivation and having parents who are more highly educated also contribute to this prediction. The adjusted R squared value was .43. This indicates that 43% of the variance in math achievement was explained by the model. According to Cohen (1988), this is a large effect.

Table 6.1

Means, Standard Deviations, and Intercorrelations for Math Achievement and Predictor Variables (N=73)

Variable	*M*	*SD*	1	2	3	4
Math Achievement	12.60	6.76	.32**	.50**	.39**	-.30**
Predictor variable						
1. Motivation scale	2.87	.64	-	.08	.09	-.21*
2. Grades in h.s.	5.68	1.59		-	.25*	.12
3. Parent's education	4.38	2.30			-	-.23*
4. Gender	.55	.50				-

*$p < .05$; **$p < .01$.

Table 6.2

Simultaneous Multiple Regression Analysis Summary for Motivation, Grades in High School, Parent's Education, and Gender Predicting Math Achievement (N = 73)

Variable	*B*	*SEB*	β
Motivation scale	2.15	.97	.20*
Grades in h.s.	1.99	.40	.47**
Parent's education	.58	.28	.20*
Gender	-3.63	1.28	-.27**
Constant	-5.44	3.61	

Note. $R^2 = .46$; $F(4,68) = 14.44$, $p < .001$
*$p < .05$; **$p < .01$.

Problem 6.3: Hierarchical Multiple Linear Regression

In Problem 6.3, we will use the **hierarchical** approach, which enters variables in a series of blocks or groups, enabling the researcher to see if each new group of variables adds anything to the prediction produced by the previous blocks of variables. This approach is an appropriate method to use when the researcher has a priori ideas about how the predictors go together to predict the dependent variable. In our example, we will enter *gender* first and then see if any of the other variables make an additional contribution. This method is intended to control for or eliminate the effects *of gender* on the prediction.

6.3. If we control for *gender* differences in *math achievement,* do any of the other variables significantly add anything to the prediction over and above what *gender* contributes?

We will include all of the variables from the previous problem; however, this time we will enter the variables in two separate blocks to see how *motivation, grades in high school, and parents' education* improve on prediction from *gender* alone.

• Click on the following: **Analyze => Regression => Linear.**

• Click on **Reset.**

• Select *math achievement* and click it over to the **Dependent** box (dependent variable).

• Next, select *gender* and move it to the over to the **Independent(s)** box (independent variables).

• Select **Enter** as your **Method.** (See Fig. 6.4.)

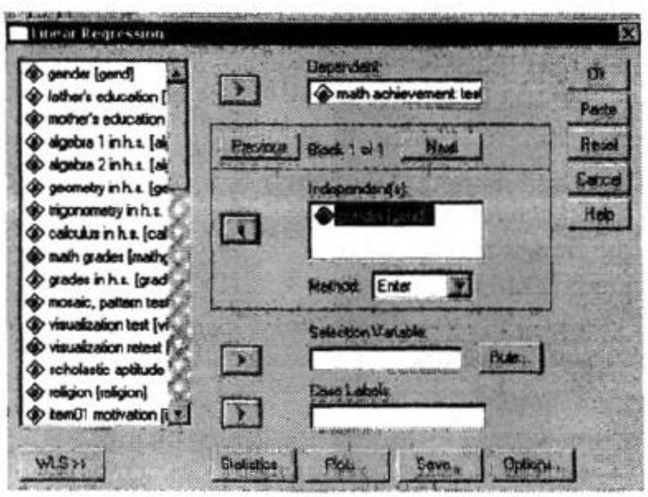


Fig. 6.4. Linear regression.

• Click on Next beside Block 1 of 1. You will notice it changes to **Block 2 of 2.**

• Then move *motivation scale, grades in h.s., and parent's education* to the **Independents) box** (independent variables). Under **Method,** select **Enter.** The window should look like Fig. 6.5.

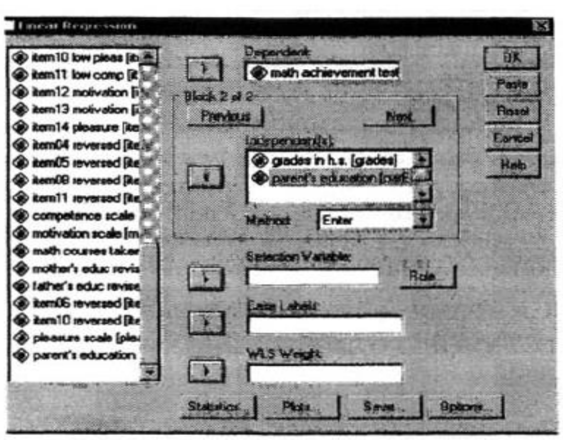


Fig. 6.5. Hierarchical Regression.

• Click on **Statistics,** click on **Estimates** (under **Regression Coefficients),** and click on **Model fit and R squared change.** (See Fig. 6.2.)

• Click on **Continue.**

• Click on OK. Compare your output and syntax to Output 6.3.

Output 6.3: Hierarchical Multiple Linear Regression

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA CHANGE
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT mathach
  /METHOD=ENTER gender  /METHOD=ENTER motivation grades parEduc  .
```

Regression

Variables Entered/Removed [b]

Model	Variables Entered	Variables Removed	Method
1	gender [a]	.	Enter
2	grades in h.s., motivation scale, parents' education [a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: math achievement test

In the first column of this table there are two models (1 and 2). This indicates that first we tested a model with *gender* as a predictor, then we added the other predictors and tested that model (Model 2).

Footnotes provide you with relevant information.

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	.303(a)	.092	.079	6.48514	.092	7.158	1	71	.009
2	.678(b)	.459	.427	5.11249	.368	15.415	3	68	.000

a Predictors: (Constant), gender

b Predictors: (Constant), gender, grades in h.s., motivation scale, parents' education

The Model Summary output shows there were two models run: Model 1 (in the first row) and Model 2 (in the second row). It also shows that the addition of *grades, motivation,* and *parents' education* significantly improved on the prediction by *gender* alone, explaining almost 37% additional variance.

ANOVA(c)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	301.026	1	301.026	7.158	.009(a)
	Residual	2986.050	71	42.057		
	Total	3287.076	72			
2	Regression	1509.723	4	377.431	14.440	.000(b)
	Residual	1777.353	68	26.138		
	Total	3287.076	72			

a Predictors: (Constant), gender

b Predictors: (Constant), gender, grades in h.s., motivation scale, parents' education

c Dependent Variable: math achievement test

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	14.838	1.129		13.144	.000
	gender	-4.080	1.525	-.303	-2.675	.009
2	(Constant)	-5.444	3.605		-1.510	.136
	gender	-3.631	1.284	-.269	-2.828	.006
	motivation scale	2.148	.972	.203	2.211	.030
	grades in h.s.	1.991	.400	.468	4.972	.000
	parents' education	.580	.280	.198	2.070	.042

a. Dependent Variable: math achievement test

Excluded Variables(b)

Model		Beta In	T	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	motivation scale	.264(a)	2.358	.021	.271	.956
	grades in h.s.	.546(a)	5.784	.000	.569	.987
	Parents' education	.343(a)	3.132	.003	.351	.949

a Predictors in the Model: (Constant), gender
b Dependent Variable: math achievement test

Interpretation of Output 6.3
We did not need to recheck the assumptions for this problem, because we checked them in Problem 6.2 with the same variables.

The **Descriptives** and **Correlations** tables would have been the same as those in Problem 6.2 if we had checked the Descriptive box in the Statistics window.

The other tables in this output are somewhat different than the previous two outputs. This difference is because we entered the variables in two steps. Therefore, this output has two models listed, Model 1 and Model 2. The information in Model 1 is for *gender* predicting *math achievement*. The information in Model 2 is *gender* plus *motivation, grades in h.s.*, and *parents' education* predicting *math achievement*.

We can see from the **ANOVA** table that when *gender* is entered by itself, it is a significant predictor of *math achievement*, $F(1,71) = 7.16$, $p = .009$; however, the model with the addition of the other predictor variables is a better model for predicting *math achievement* $F(4,68) = 14.44, p < .001$. That Model 2 is better than Model 1 can also be seen in the **Model Summary** table by the increase in the adjusted R^2 value from $R^2 = .08$ to an $R^2 = .43$, $F(3, 83) = 15.42, p < .001$.

Note also that results for the final model, with all of the variables entered, is identical to the model in Problem 6.2. No matter how they are entered, the same regression coefficients are selected to produce the best model with the same set of predictors and the same dependent variable.