6: Regression and Multiple Regression

Objectives

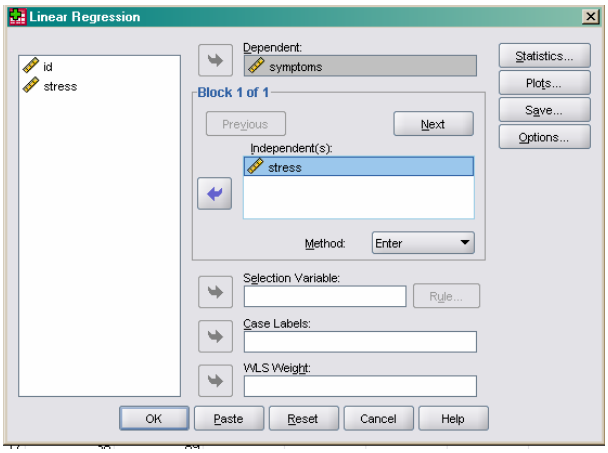
- Calculate regressions with one independent variable
- Calculate regressions with multiple independent variables
- Scatterplot of predicted and actual values
- Calculating residuals and predicted values

Regression

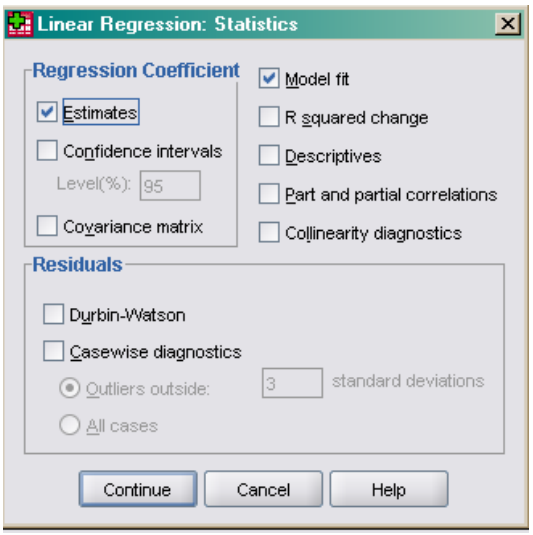
Regression allows you to predict variables based on another variable. In this chapter we will focus on linear regression or relationships that are linear (a line) rather than curvilinear (a curve) in nature. Let's begin with the example used in the text in which mental health symptoms are predicted from stress.

✔ **Open** *symptoms and stress.sav*.

✔ Select **Analyze/Regression/Linear**.



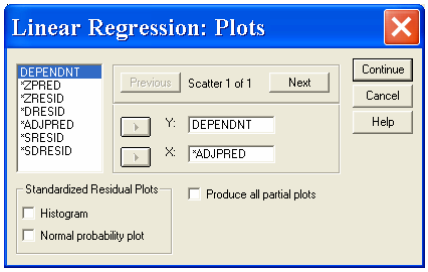
✔ Select symptoms as the **Dependent** variable and stress as the **Independent** variable. Then, click on **Statistics** to explore our options. The following dialog box will appear.



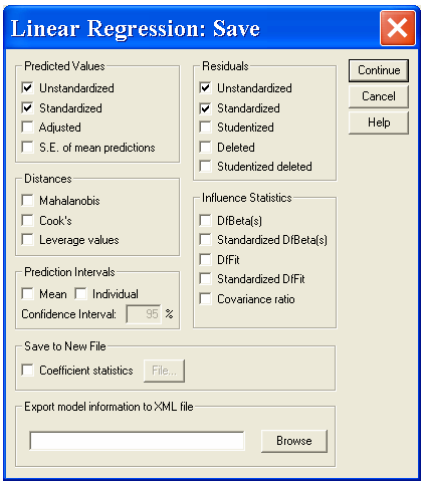
- As you can see there are many options. We will focus only on information covered in the textbook. **Estimates** and **Model Fit** are selected by default. Leave them that way. Then select **Descriptives** and **Part and partial correlations**. SPSS will then calculate the mean and standard deviation for each variable in the equation and the correlation between the two variables. Then, click **Continue**.

- At the main dialog box, click on **Plots** so we can see our options.

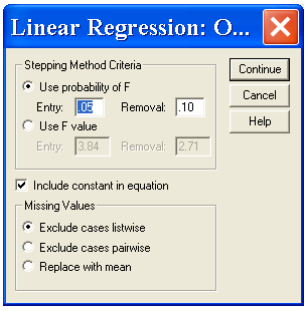
- It looks like we can create scatterplots here. Click **Help** to see what the abbreviations represent. I'd like to plot the Dependent variable against the predicted values to see how close they are. Select **Dependnt** for **Y** and **Adjpred** for **X**. Adjpred is the adjusted prediction. Used **Help/Topics/Index** to find out what this means for yourself. Then, click **Continue**.



- In the main dialog box, click **Save**, and the dialog box to the left will appear. For **Predicted Values**, select **Unstandardized** and **Standardized**. For **Residuals**, also select **Unstandardized** and **Standardized**. Now, SPSS will save the predicted values of symptoms based on the regression equation and the residual or difference between the predicted values and actual values of symptoms in the data file. This is a nice feature. Remember, the standardized values are based on z score transformations of the data whereas the unstandardized values are based on the raw data. Click **Continue**.



✔ Finally, click on **Options**.



✔ **Including a constant in the equation** is selected by default. This simply means that you want both a slope and an intercept (the constant). That's good. We will always leave this checked. Excluding cases listwise is also fine. We do not have any missing cases in this example

Descriptive Statistics

	Mean	Std. Deviation	N
SYMPTOMS	90.70	20.27	107
STRESS	21.47	13.10	107

Correlations

		SYMPTOMS	STRESS
Pearson Correlation	SYMPTOMS	1.000	.506
	STRESS	.506	1.000
Sig. (1-tailed)	SYMPTOMS	.	.000
	STRESS	.000	.
N	SYMPTOMS	107	107
	STRESS	107	107

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	STRESS[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: SYMPTOMS

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.506[a]	.256	.249	17.56

a. Predictors: (Constant), STRESS

b. Dependent Variable: SYMPTOMS

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11148.382	1	11148.382	36.145	.000[a]
	Residual	32386.048	105	308.439		
	Total	43534.430	106			

a. Predictors: (Constant), STRESS

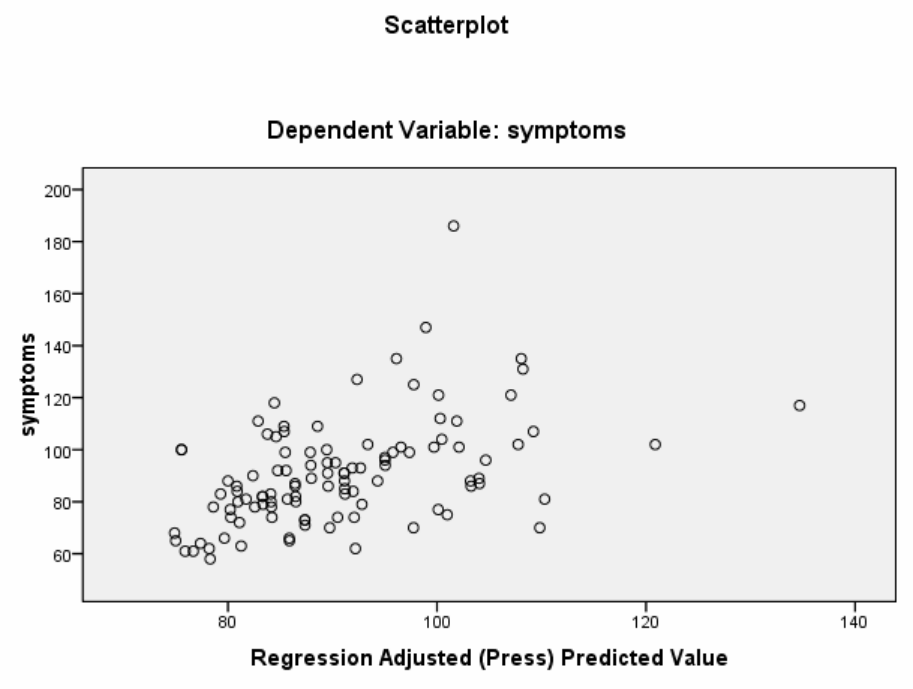
b. Dependent Variable: SYMPTOMS

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	73.890	3.271		22.587	.000			
	STRESS	.783	.130	.506	6.012	.000	.506	.506	.506

a. Dependent Variable: SYMPTOMS

Charts



- How does our output compare to the output presented in the textbook? Take a moment to identify all of the key pieces of information. Find r^2, find the ANOVA used to test the significance of the model, find the regression coefficients used to calculate the regression equation. One difference is that the text did not include the scatterplot. What do you think of the scatterplot? Does it help you see that predicting symptoms based on stress is a pretty good estimate? You could add a line of best fit to the scatterplot using what you learned in Chapter 5.

- Now, click **Window/Symptoms and stress.sav** and look at the new data (residuals and predicted values) in your file. A small sample is below. Note how they are named and labeled.

	id	stress	symptoms	pre_1	res_1	zpr_1	zre_1
1	1	30	99	97.3830	Unstandardized Predicted Value		09207
2	2	27	94	95.03368	-1.03368	.42248	-.05886
3	3	9	80	80.93762	-.93762	-.95202	-.05339
4	4	20	70	89.55188	-19.5519	-.11204	-1.11328

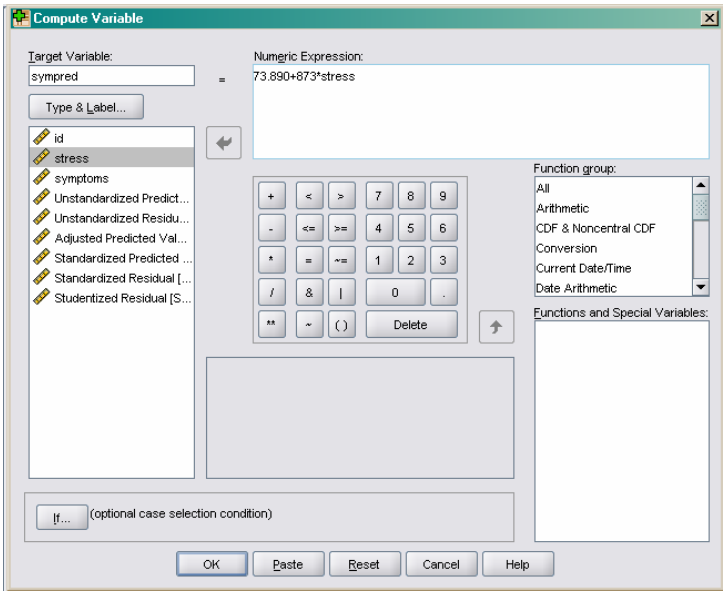
Let's use what we know about the regression equation to check the accuracy of the scores created by SPSS. We will focus on the unstandardized predicted and residual values. This is also a great opportunity to learn how to use the Transform menus to perform calculations based on existing data.

We know from the regression equation that:
Symptoms Predicted or $\hat{Y}$ = 73.890 + .783* Stress.

We also know that the residual can be computed as follows:
Residual = Y-$\hat{Y}$ or Symptoms – Symptoms Predicted Values.

We'll use SPSS to calculate these values and then compare them to the values computed by SPSS.

✔ In the Data Editor window, select **Transform/Compute**.



- Check the Data Editor to see if your new variable is there, and compare it to pre_1. Are they the same? The only difference I see is that our variable is only expressed to 2 decimal places. But, the values agree.

- Follow similar steps to calculate the residual. Click on **Transform/Compute**. Name your **Target Variable** sympres and **Label** it symptoms residual. Put the formula symptoms-sympred in the **Numeric Expression** box by double clicking the two pre-existing variables and typing a minus sign between them. Then, click **Ok**.

- Compare these values to res_1. Again they agree. A portion of the new data file is below.

	id	stress	symptoms	pre_1	res_1	zpr_1	zre_1	sympred	sympres
1	1	30	99	97.38302	1.61698	.65157	.09207	9	Symptoms Predicted
2	2	27	94	95.03368	-1.03368	.42248	-.05886	95.03	-1.03
3	3	9	80	80.93762	-.93762	-.95202	-.05339	80.94	-.94
4	4	20	70	89.55188	-19.5519	-.11204	-1.11328	89.55	-19.55

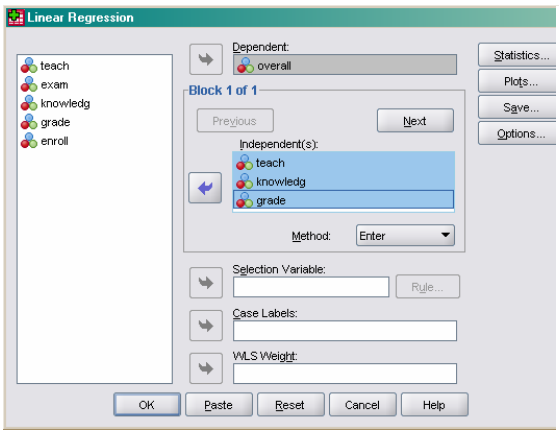
Now that you are confident that the predicted and residual values computed by SPSS are exactly what you intended, you won't ever need to calculate them yourself again. You can simply rely on the values computed by SPSS through the Save command.

Multiple Regression

Now, let's move on to multiple regression. We will predict the dependent variable from multiple independent variables. This time we will use the course evaluation data to predict the overall rating of lectures based on ratings of teaching skills, instructor's knowledge of the material, and expected grade.

- **Open** *course evaluation.sav*. You may want to save *symptoms and stress.sav* to include the residuals. That's up to you.

- Select **Analyze/Regression/Linear**.



- Select overall as the **Dependent** variable, and teach, knowledge, and grade as the **Independents**. Since there are multiple independent variables, we need to think about the Method of entry. As noted in the text, stepwise procedures are seductive, so we want to select **Enter** meaning all of the predictors will be entered simultaneously.

- Click **Statistics** and select **Descriptives** and **Part and partial correlations**. Click **Continue**.
- Click **Plots** and select **Dependnt** as **Y** and **Adjpred** as **X**. Click **Continue**.
- Click **Save** and select the **Residuals** and **Predicted** values of your choice. Click **Continue**.
- Click **Ok** at the main dialog box. The output follows.

Descriptive Statistics

	Mean	Std. Deviation	N
OVERALL	3.55	.61	50
TEACH	3.66	.53	50
KNOWLEDG	4.18	.41	50
GRADE	3.49	.35	50

Correlations

		OVERALL	TEACH	KNOWLEDG	GRADE
Pearson Correlation	OVERALL	1.000	.804	.682	.301
	TEACH	.804	1.000	.526	.469
	KNOWLEDG	.682	.526	1.000	.224
	GRADE	.301	.469	.224	1.000
Sig. (1-tailed)	OVERALL	.	.000	.000	.017
	TEACH	.000	.	.000	.000
	KNOWLEDG	.000	.000	.	.059
	GRADE	.017	.000	.059	.
N	OVERALL	50	50	50	50
	TEACH	50	50	50	50
	KNOWLEDG	50	50	50	50
	GRADE	50	50	50	50

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	GRADE, KNOWLEDG, TEACH[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: OVERALL

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.863[a]	.745	.728	.32

a. Predictors: (Constant), GRADE, KNOWLEDG, TEACH

b. Dependent Variable: OVERALL

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	13.737	3	4.579	44.741	.000[a]
	Residual	4.708	46	.102		
	Total	18.445	49			

a. Predictors: (Constant), GRADE, KNOWLEDG, TEACH

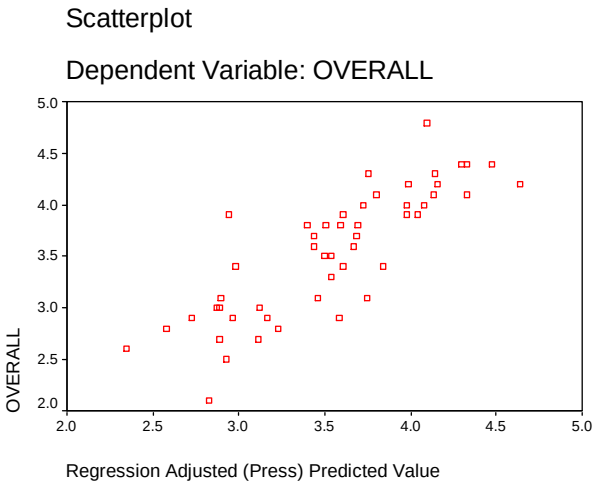
b. Dependent Variable: OVERALL

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.927	.596		-1.556	.127
	TEACH	.759	.112	.658	6.804	.000
	KNOWLEDG	.534	.132	.355	4.052	.000
	GRADE	-.153	.147	-.088	-1.037	.305

a. Dependent Variable: OVERALL

Charts



- Compare this output to the results in the text. Notice the values are the same, but the styles are different since the output in the book (earlier edition) is from Minitab, a different data analysis program.

- **Exit** SPSS. It's up to you to decide if you want to save the changes to the data file and the output file.

In this chapter, you have learned to use SPSS to calculate simple and multiple regressions. You have also learned how to use built in menus to calculate descriptives, residuals and predicted values, and to create various scatterplots. As you can see, SPSS

has really simplified the process. Complete the following exercises to increase your comfort and familiarity with all of the options.

Exercises

1. Using data in *course evaluations.sav*, predict overall quality from expected grade.

2. To increase your comfort with Transform, calculate the predicted overall score based on the regression equation from the previous exercise. Then calculate the residual. Did you encounter any problems?

3. Using data in *HeightWeight.sav*, predict weight from height and gender. Compare your results to the output in Table 11.6 of the textbook.

4. Using the data in *cancer patients.sav*, predict distress at time 2 from distress at time 1, blame person, and blame behavior. Compare your output to the results presented in Table 11.7 in the textbook.