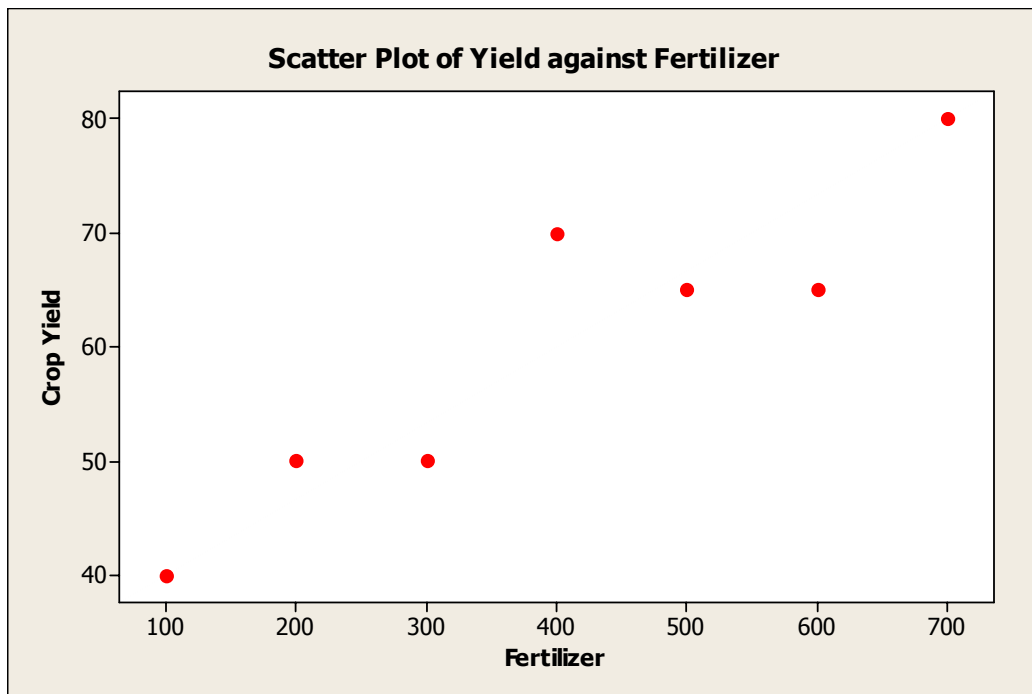


Using Minitab for Regression Analysis: An extended example

The following example uses data from another text on fertilizer application and crop yield, and is intended to show how Minitab can be used to generate the statistical measures discussed in *Anderson Sweeney and Williams*. The data here have already been input, and stored in C1 and C2. The scatter plot below is similar to A, S, & W Figure 14.3 on page 559. You can find additional discussion of scatter plots in the *Minitab Handbook* on pages 158-63.

This is formed using GRAPH > SCATTERPLOT > SIMPLE



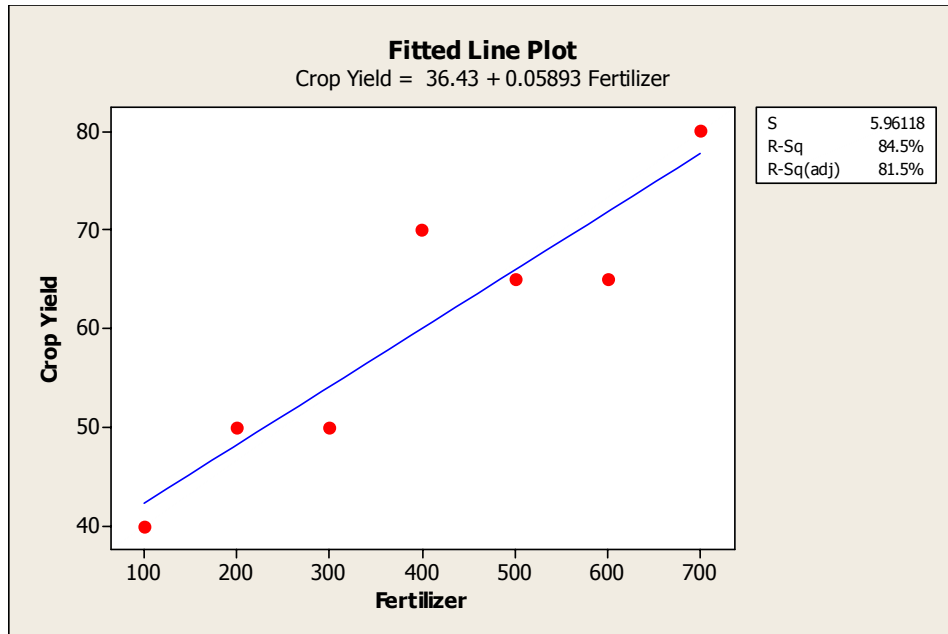
There seems to be a strong positive relationship between these two variables. One way to measure the strength of the relationship is with the correlation coefficient. For discussion see A, S & W, pages 110-12, 572-3 or the *Minitab Handbook*, pages 296-97. In Minitab 14 you can compute a correlation by clicking on STAT > BASIC STATISTICS > CORRELATION. Minitab uses formula (3.12) on page 110 of A, S & W to do the calculation.

```
MTB > Correlation 'Fertilizer' 'Crop Yield'.
```

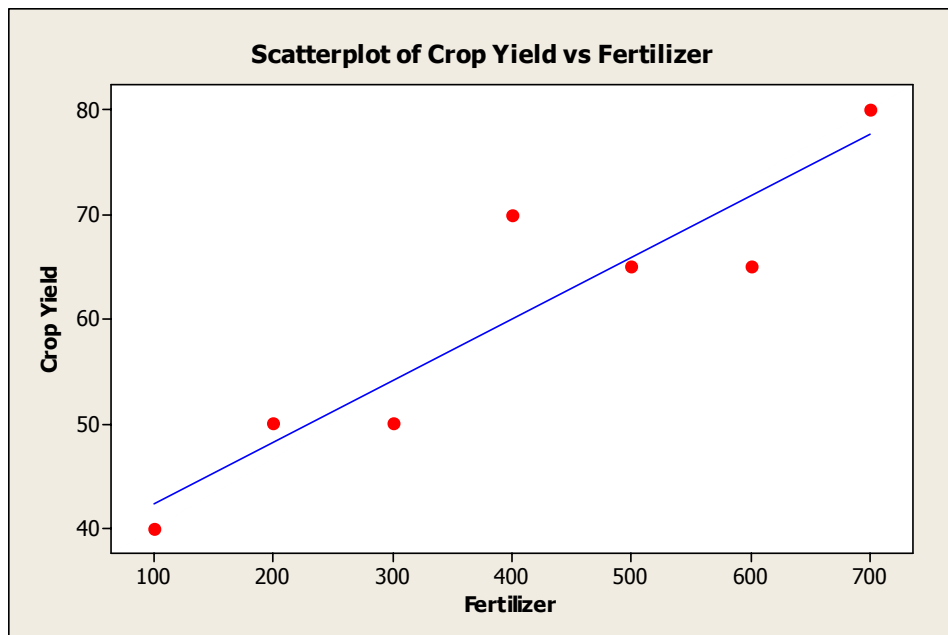
Correlations: Fertilizer, Crop Yield

```
Pearson correlation of Fertilizer and Crop Yield = 0.920  
P-Value = 0.003
```

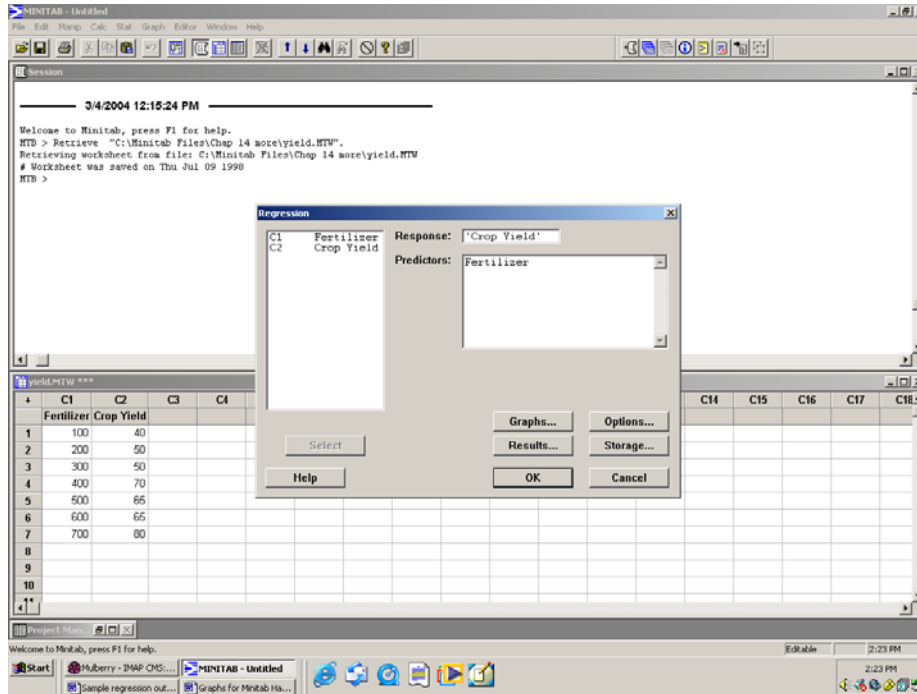
We can overlay a fitted line on our scatter plot. We can obtain the plot either with or without the estimated regression equation. To produce a plot with the estimated equation, use STAT > REGRESSION > FITTED LINE PLOT, and selecting the Linear alternative (See Minitab Handbook, pp. 320-21).



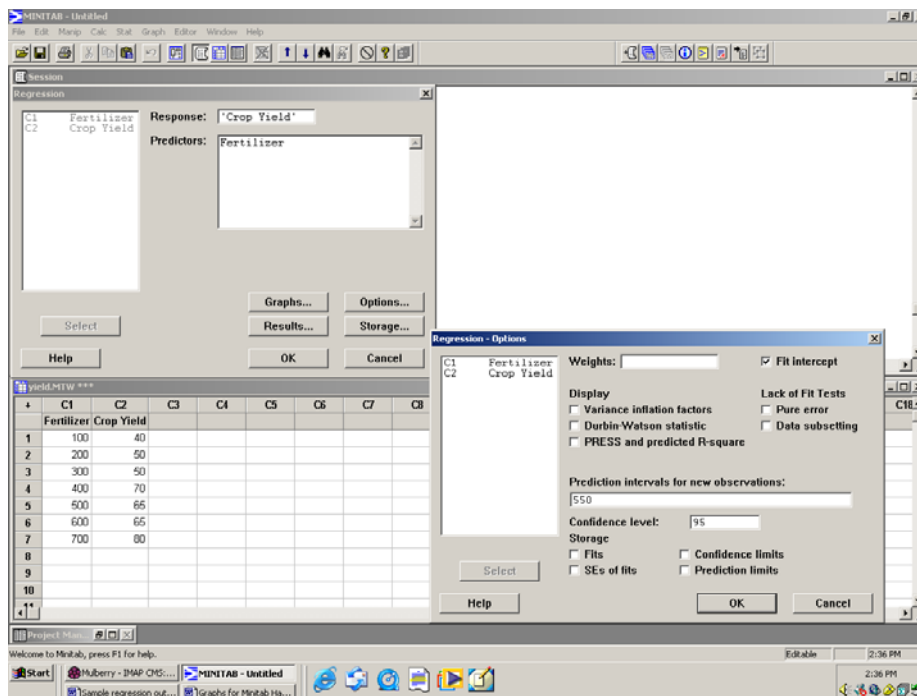
Or this simpler picture from GRAPH > SCATTERPLOT > WITH REGRESSION



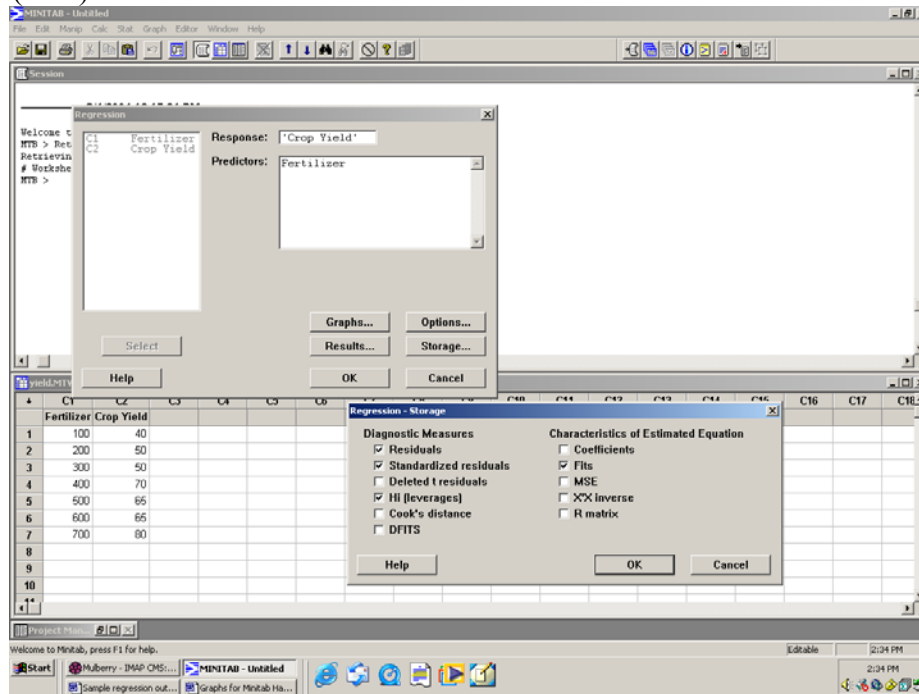
To estimate a regression line, the command in Minitab is “Regress,” which we access by clicking on STAT > REGRESSION > REGRESSION. We fill it in as shown, with the Y variable entered into “Response:” and the X variable entered into “Predictors:”



To predict, we click on the box labeled Options... and insert the X value (the value of Fertilizer) for which we would like to predict. Here, I chose to predict for X=550.



By clicking on the box labeled Storage..., I can ask Minitab to store some information for me to use later: the residuals, the standardized residuals, the leverages, and the fitted values. If you are uncertain about what these terms mean, you can consult ASW. Residuals are defined in ASW by equation (14.28); standardized residuals by equation (14.32); leverages by equation (14.33); and fitted values the values of \hat{y} defined by equation (14.3).



When you click on OK > OK and Minitab executes the command, the following code appears in the session window. If you examine it you will see that Minitab begins by labeling the columns in which it will store the residuals (“RESI1”), the standardized residuals (“SRES1”), the leverages (“HI1”), and the fitted values (“FITS1”). The main command, the “Regress” command, instructs Minitab to regress the Y variable ‘Crop Yield’ on one X variable, named ‘Fertilizer.’ Then Minitab uses various subcommands to store the residuals, standardized residuals, leverages, and fitted values. The subcommand “predict” is asking for a prediction when X=550, and the final subcommand asks Minitab to suppress some optional output.

Here is the computer output.

```
MTB > Name c3 "RESI1" c4 "SRES1" c5 "HI1" c6 "FITS1"
MTB > Regress 'Crop Yield' 1 'Fertilizer';
SUBC> Residuals 'RESI1';
SUBC> SResiduals 'SRES1';
SUBC> Hi 'HI1';
SUBC> Fits 'FITS1';
SUBC> Constant;
SUBC> Predict 550;
SUBC> Brief 2.
```

Below the code you find your results:

Regression Analysis: Crop Yield versus Fertilizer

The regression equation is
Crop Yield = 36.4 + 0.0589 Fertilizer

Predictor	Coef	SE Coef	T	P
Constant	36.429	5.038	7.23	0.001
Fertilizer	0.05893	0.01127	5.23	0.003

S = 5.96118 R-Sq = 84.5% R-Sq(adj) = 81.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	972.32	972.32	27.36	0.003
Residual Error	5	177.68	35.54		
Total	6	1150.00			

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	68.84	2.82	(61.60, 76.08)	(51.89, 85.79)

Values of Predictors for New Observations

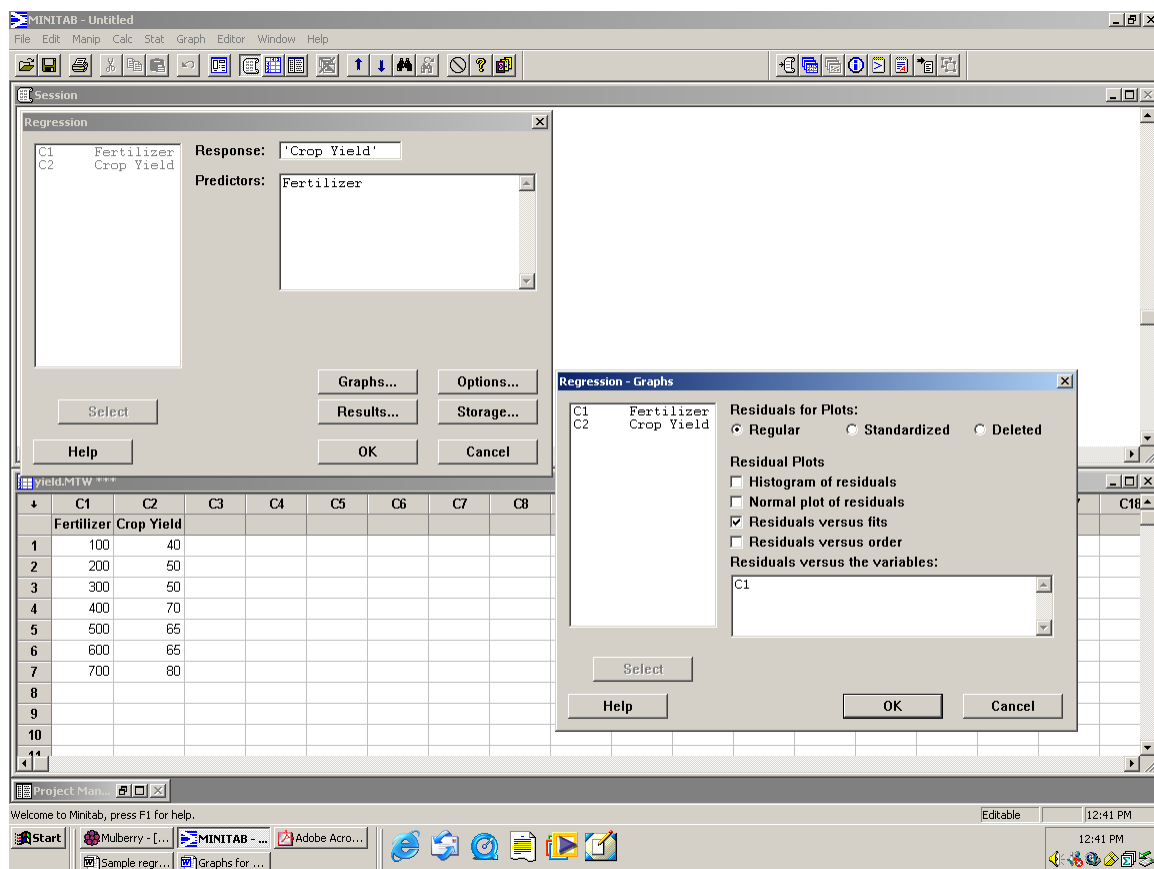
New Obs	Fertilizer
1	550

The estimated intercept, b_0 , computed using formula (14.6), is the number 36.429 found immediately under the column heading Coef. The estimated slope, b_1 , computed using formula (14.7), is .05893, also under the column heading Coef. The estimated line is rewritten for your convenience at the top of the output. For further discussion, you may wish to consult page 560 of A, S, & W or pp. 297-306 of the *Minitab Handbook*. The Sum of Squares due to error, formula (14.8) of the text, is the number 177.68 in the column headed "SS." The total Sum of Squares, formula (14.9), is the number 1150.00 immediately below it. Finally, the Sum of Squares Due to the Regression, formula (14.10), is 972.32 at the top of the column. The Coefficient of Determination, r^2 , given by formula (14.12), is printed in the middle of the output, as R-Sq =84.5%. Note that $SSR/SST = 972.32/1150 = .845$, which is where R-Sq comes from. The Mean Squared Error, or s^2 given by formula (14.15), is under the heading MS. It is $35.54 = 177.68/5$ and its square root, the standard error of the estimate, formula (14.16), is $s=5.961$. The estimated standard deviation of b_1 , given by formula (14.18), is in the column headed by "SE Coef" and is equal to 0.01127. Just above this is the estimated standard deviation of b_0 , a formula missing from ASW, but equal in this instance to 5.038. The test statistic for b_1 , given by formula (14.19) is in the column headed by "t-ratio" and is equal to 5.23.

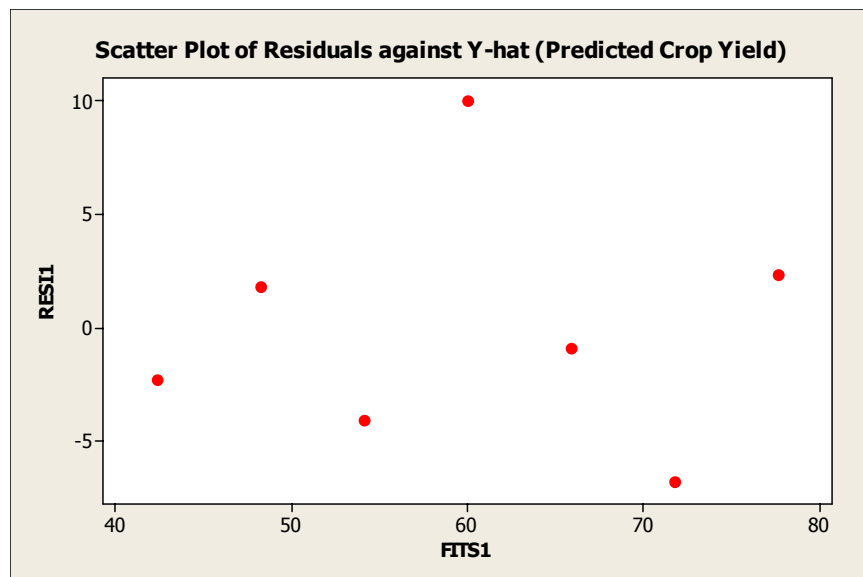
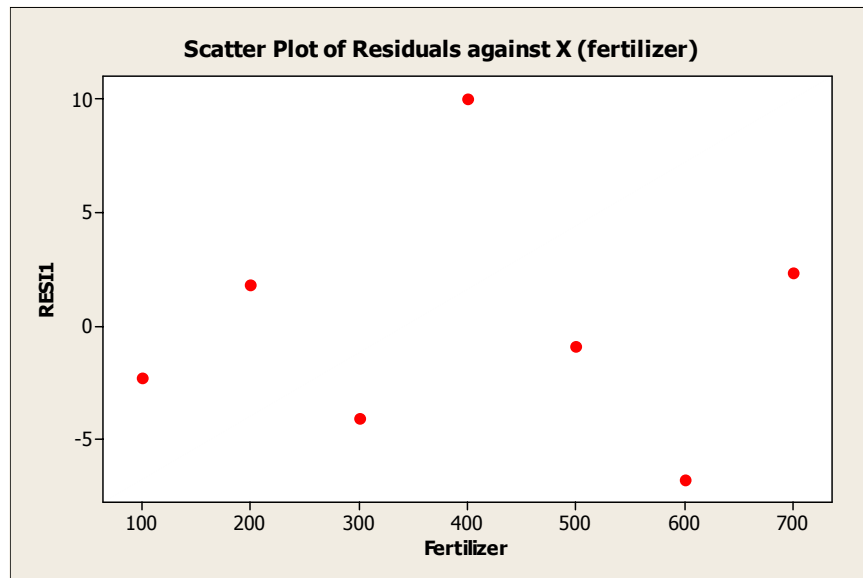
The corresponding test statistic for b_0 , another formula missing from ASW, is just above that of b_1 and is equal to 7.23. The F statistic given by formula (14.21) is given on the output, under the heading “F” and is equal to $27.36 = 972.32/35.54$.

Finally, the last thing on the output is a prediction generated by the “predict” subcommand, which requested a prediction for a value of fertilizer = 550. The predicted level of crop yield is 68.84, and the Confidence interval Estimate of $E(Y_p)$, computed using formulas (14.23, 14.24), is given by the “95% C.I.” and is (61.60, 76.08). The Prediction Interval Estimate of Y_p , computed using formulas (14.26, 14.27), is given by the “95% P.I.” and is (51.89, 85.79).

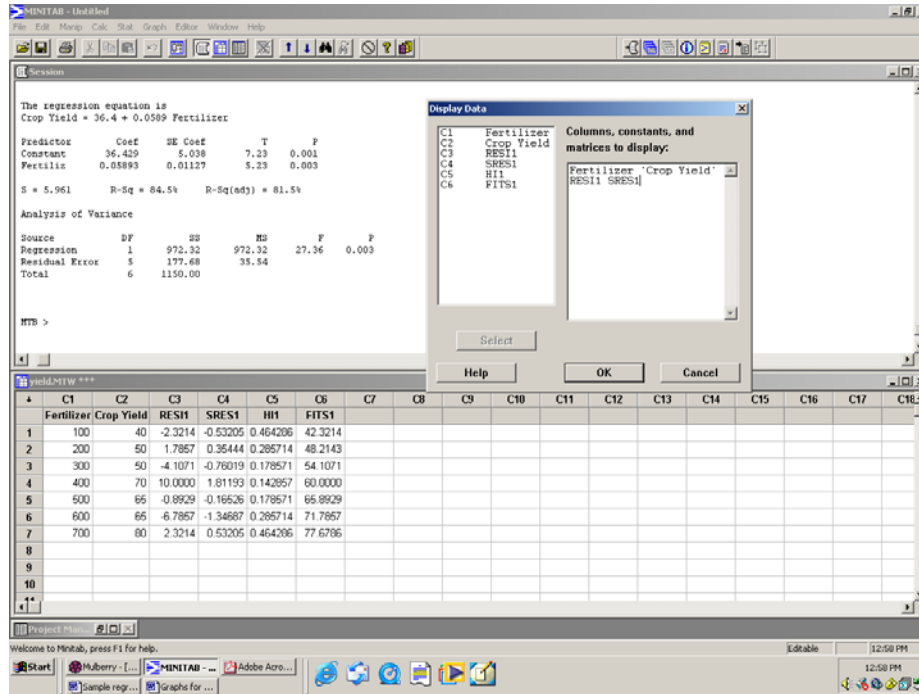
Figures 14.11, 14.12, and 14.13, pp. 600-2 of A, S & W, show how to examine the assumptions of the model by examining plots of the residuals. In the following discussion, I will show how to make the graphs from scratch. However, this is not really necessary, because the process can be automated. As an example of the automated procedure, click on Graphs... and then fill in the following information to get plots of residuals against both fitted values and Fertilizer.



Of course, once the residuals, etc., are stored you can equally well create graphs of them yourself, using `GRAPH > SCATTERPLOT > SIMPLE`. These two plots, showing residuals plotted against X and against fitted values, are just like Figures 14.11 and 14.13 in the book. They are displayed on the same page so that you will notice that they are identical, except for the units on the X-axis, just as 14.11 and 14.13 are. These plots differ when there is more than one explanatory variable – that is, in multiple regression. However, when there is only one explanatory variable – simple regression – predicted y is just a linear transformation of x and the graphs are redundant.



Standardized residuals, computed using equation (14.32), are also of interest. Once standardized residuals and raw residuals have been calculated by Minitab, you can print them out in the session window using the following command found under MANIP > DISPLAY DATA.

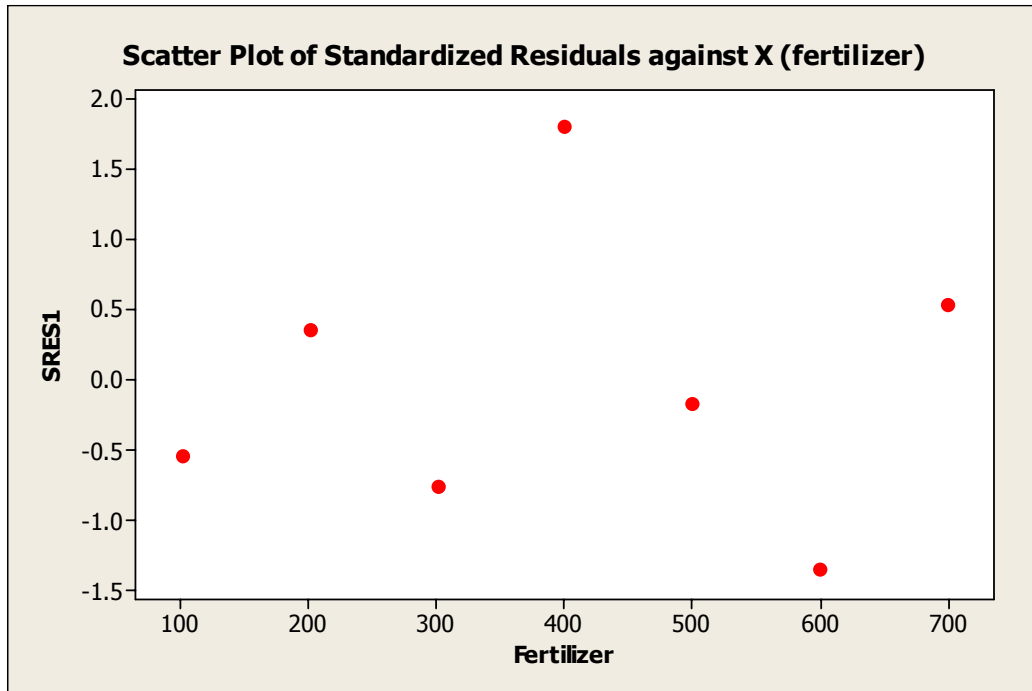


This gives rise to the following output.

Data Display

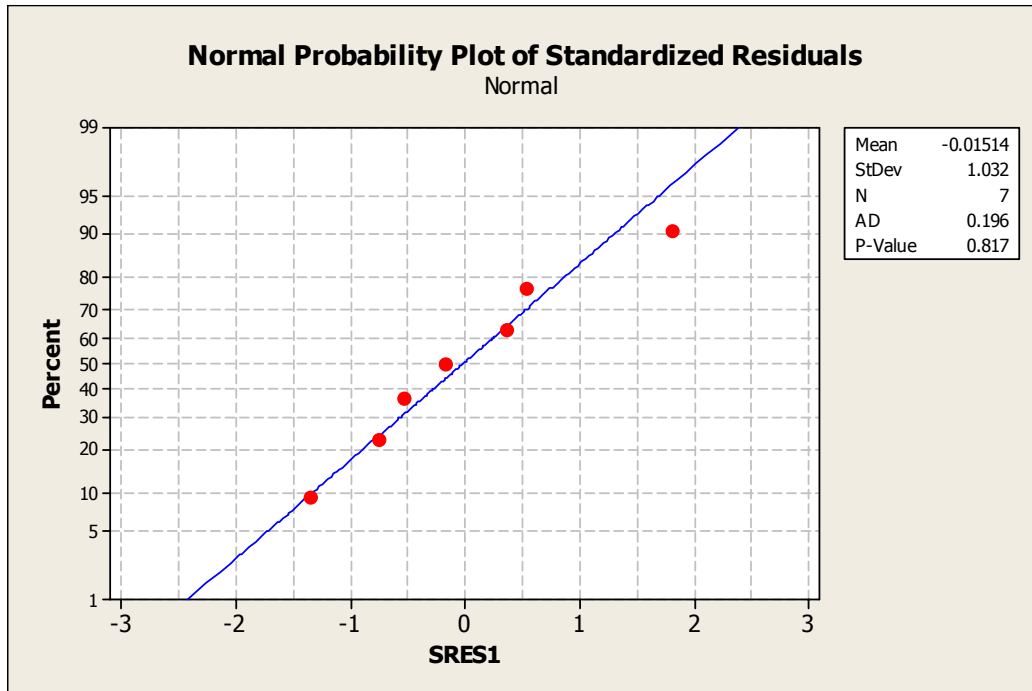
Row	Fertilizer	Crop Yield	RES11	SRES1
1	100	40	-2.3214	-0.53205
2	200	50	1.7857	0.35444
3	300	50	-4.1071	-0.76019
4	400	70	10.0000	1.81193
5	500	65	-0.8929	-0.16526
6	600	65	-6.7857	-1.34687
7	700	80	2.3214	0.53205

Standardized residuals can then be plotted and examined for evidence that regression assumptions are being violated, as in Figure 14.14 of ASW.



Finally, in Figure (14.15) of A, S & W, the authors show a normal probability plot for the standardized residuals. This is an attempt to check whether the errors are plausibly normally distributed. Here is the corresponding plot. The command "nscore c3 c8" computes the normal score of the standardized residuals in c3, and stores the result in c8, like Table 14.9 of A, S & W. For a description of the command, see the *Minitab Handbook*, pp. 240-2. In the latest version of Minitab, you can find the command in the menu by looking under STAT > BASIC STATISTICS > NORMALITY TEST.

Since the plot below is approximately linear, the errors appear to be (approximately, at least) normally distributed.



One way of detecting influential observations is to compute the leverage of the observations. This is discussed in the text on pages 609-11, and mentioned in the *Minitab Handbook*, p. 400. Below we display the leverage of each observation. These were computed using formula (14.33) in A, S & W. Minitab considers leverage noteworthy if it is greater than $3p/n$, where p is the number of coefficients, including the constant. In this case $3p/n = 6/7 = .857$, so none are noteworthy.

```
MTB > print c1 c5
```

Data Display

Row	Fertilizer	HI1
1	100	0.464286
2	200	0.285714
3	300	0.178571
4	400	0.142857
5	500	0.178571
6	600	0.285714
7	700	0.464286